*Mapping
the Way to
a Better Future ...*

Plant Genome Data and Information Center

1991-1992

Plant Genome Data and Information Center
1991 - 1992

Prepared and Compiled By:
Susan McCarthy, Ph.D., Coordinator
December 11, 1992

For More Information Contact:
Plant Genome Data and Information Center
10301 Baltimore Blvd., Rm 1402
Beltsville, MD 20705

Acknowledgements

I would like to acknowledge and thank the following contributors and reviewers: Keith Russell, Pamela Andre, Sarah Thomas, Leslie Kulp, Rose Broome, Lori Starr, Idalia Acosta, Shirley Edwards, Caroline Early, Barbara Buchanan. A special thanks is given to Terrance Henrichs for her very able technical assistance.

Preface

This document is a compendium of information that identifies the activities of the Plant Genome Data and Information Center (PGDIC) at the National Agricultural Library (NAL). The unique organizational structure of this Center reflects the diverse activities through which NAL contributes to the USDA Plant Genome Research Program. Key staff members administratively housed in the three divisions of the library, Public Services (PSD), Technical Services (TSD) and Information Systems(ISD), provide expertise to accomplish the work of the Plant Genome Data and Information Center. **PGDIC Personnel** identifies the key staff members by division. The rest of this compendium is organized according to the services listed in **PGDIC's Divisional Services**, which categorizes the activities of the Center and identifies the involvement of each division in these activities.

The programs of the Plant Genome Data and Information Center are focused in six key areas of activities. **Information Services** are designed to respond to specific information needs of requestors through reference and referral services, to provide administrative and technical support to the Plant Genome Research Program, and to produce and disseminate information products such as Probe, the official publication of the Program. As a result of **Outreach** activities, targeted populations are informed about the overall Plant Genome Research Program and the specific information services available through PGDIC. Additionally, the Center is providing training in workshops that improve search skills of scientist and librarians in databases relevant to the Program.

The collection at NAL is a primary resource for USDA researchers and others who need access to information relevant to plant genome mapping. **Collection Development** activities at NAL identify, acquire, and process journals, books, and information in other formats so that it can be made available directly to USDA and, through the interlibrary loan network, to the rest of the world. The **AGRICOLA** database, the bibliographic database produced at NAL, is the premier vehicle through which persons interested in agricultural information can identify significant literature to meet their information needs. Activities to enhance **AGRICOLA** are instrumental in improving access to published literature relevant to plant genome mapping.

Database Development activities are targeted toward the production of the Plant Genome Database, which will become the mechanism for accessing map data. Through **National and International Networking Activities**, PGDIC has established contacts and communication links to other significant mapping efforts and related professional societies.

This compendium serves as the background material documenting the activities of the National Agricultural Library in the Plant Genome Mapping Program.

Table of Contents

Introduction

Today's population is expected to double by the year 2030. At the same time, agricultural producers will have to yield more substance on smaller acreage using fewer chemicals. Facing these significant challenges will require a Federal commitment to research. The U.S. Department of Agriculture's Plant Genome Research Program represents such a commitment.

The USDA Plant Genome Research Program is a multi-agency effort, with the Agricultural Research Service serving as the lead agency. The Cooperative State Research Service handles the competitive grants as part of the National Research Initiative (NRI) Competitive Grants Program. This program supports research applied to developing chromosomal maps of the commercial crop species, as well as grants researching the isolation and manipulation of economically important genes. Additionally, the NRI is funding grants aimed at developing new and innovative technologies to assist with long term Program goals.

The Plant Genome Research Program is committed to serving the broad information needs of the research community. Databases to help organize, analyze and store the Program research results are under development. A central database which will store information from all sources will be housed at the National Agricultural Library. The data from the central database will be widely available to our user community. Data will initially come from several species groups which have been funded through the Agricultural Research Service to collect, evaluate and digitize the data most important for their user groups. To this end, the species groups have developed their own independent databases which are also available to our user community.

A comparative database will push crop improvement practices into the 21st century. Map based breeding will allow breeders to broaden the genetic base of crops by breeding with wild relatives. Current breeding practices do not rely on these valuable genetic resources, since too many undesirable traits are carried into the breeding populations, a phenomenon known as linkage drag. With the advent of map based breeding, complementary breeding partners will be developed that can virtually eliminate linkage drag.

Map based breeding also will serve to shorten breeding cycles. In point of fact, map based breeding projects have been initiated for species having long generation cycles. Notably, apple breeders who normally must wait six years to evaluate the progeny of a mating cross may now learn the results within months. Pine breeders have developed techniques for isolating nucleic acids from the residual material in seed coats. Analyses can now describe in molecular terms the result of each and every progeny.

Recent research results also indicate that chromosomal mapping projects will help maximize research investment effectiveness. There are over 71 different crop species grown commercially in the United States. It would not be economically feasible to develop fine maps for each species, rather, with the availability of comparative mapping databases it may be possible to project research findings between species. Dr. Steven Tanksley, Cornell University, reported recently that maize and rice may share as much as 71% co-linearity of gene order. Other groups of closely related species have shown similar results.

Between 1930 and 1980, there has been a six-fold increase in corn yields. Despite the advent of chemical fertilizers and pesticides which came into widespread use in the early 1950s, it is estimated that between 33 and 80% of these yield gains can be directly attributed to improved genetic stocks. Such yield increases are becoming significantly more expensive to produce using conventional means.

Specialty crops are nearing market. These crops include varieties bred to tolerate insect pests by incorporating the *Bacillus thuringiensis* toxin protein, and viral coat proteins. Crops have been bred to be resistant to a variety of herbicides. In the area of food processing, tomatoes have been bred for higher sugar content and better handling characteristics such as bruise resistance and delayed ripening. Canola has been bred for altered lipid profiles, and the AIDS drug Compound Q has been bred into tobacco plants.

The USDA Plant Genome Research Program has an important role to play in these positive developments. It is with great excitement that we look into the future and gain appreciation for our essential role.

Plant Genome Database Activities in ISD

Pamela Andre
Associate Director for Information Systems Division
National Agricultural Library

The Information Systems Division has responsibility for supporting database development activities for the Plant Genome Data and Information Center. During 1991, the database management team has been analyzing the system requirements for the Agricultural Genome Database System. In exploring current database activities in the genetics area, NAL staff have been working with such diverse centers as the National Center for Biotechnology Information, the European Molecular Biology Laboratory, Jackson Labs, and Dupont.

The analysis phase culminated with a meeting of the Technical Advisory Committee meeting in July 1991. This meeting included representatives from key disciplines cooperating in the USDA plant genetics program, as well as experts in computer science. At that meeting, it was determined that NAL would take a lead role in coordinating design activities for the USDA cooperators.

During the design, NAL will work with USDA cooperators to develop a "generic" plant genome database design. The database will be designed to run on UNIX systems using the Sybase database management system. Work is now underway on the database design, and a prototype database is scheduled to be completed in the Spring of 1992.

Milestones in the Plant Genome Research Program

1987
- NIH/DOE Human Genome Project established

1988
- Plant Genome project proposed by J. Miksche
- Asst. Sec'y Bentley endorses ARS to lead Plant Genome Project (10/88)
- Crop & Forest Genome Mapping Conference held in Washington, D.C. (12/88)

1989
- J. Miksche appointed Director of USDA Plant Genome Office (4/89)
- Interagency Plant Genome Coordinating Committee meets (5/89 & 8/89)

1990
- ARS given $99,000 "seed money" for Plant Genome planning activities
- J. Miksche gave many presentations to: agencies, Congress, and outside groups
- S. Heller detailed to project and given responsibility for Plant Genome informatics activities
- NAL initiates the Plant Genome Data and Information Center. S. McCarthy hired as Coordinator for the Center.

1991
- ARS receives $3.674 Million additional funds for Plant Genome Project
- CSRS/NRI receives $11.0 Million for Plant Genome mapping activities
- Analysis of genomic research at ARS, Land Grant Schools, industry, and foreign groups
- Plant Genome Data and Information Center operations begin at NAL.
  Activities: Database, AGRICOLA enhancements, Collections, Newsletter, Reference and Information Services, Books
- Database Manager, D. Bigwood, hired by NAL (2/91)
- Funding for mapping & data collection/evaluation activities dispersed to ARS/FS labs (1-3/91)
- Database analysis and initial system design (2-6/91)
- NAL Technical Committee meeting for initial system/database design (7/91) (System operation planned for 7/93)
- ***Probe***, the official newsletter for the USDA Plant Genome Program begins production. (10/91)

1992
- ARS receives $3.773 Million for Plant Genome Project
- CSRS/NRI receives $13.0 Million for Plant Genome mapping activities. RFP for Plant Genome activities
- S. Heller hired (5/92)
- First International Plant Genome Conference with 415 attendees (11/92)
- Initiate examination of gene traits ready for delivery
- Database demonstrations undertaken (8-11/92)

1993

- Planned public release of operational NAL Plant Genome Database (PGD) (5/94)
- Start coordination of data analysis software (e.g. mapping - 4/93)
- Start development of additional versions of PGD (e.g. CD-ROM, tape - 10/93)
- Initiate plans to add additional species to PGD (e.g., pea, rice, lettuce, tomato - 10/93)

1994

- Second International Plant Genome Conference (PG II) to be held 1/94
- Initiate plans for satellite nodes of PGD with groups in Europe (EEC) and Asia (1/94
- Initiate plans and design to link PGD and GRIN systems (4/94)
- Satellite nodes to EEC and Japan under test (10/94)
- PGD/GRIN link under test (10/94)
- Follow analysis of gene product development

1995

- Satellite nodes and PGD/GRIN link operational (12/95)
- PGD has 7 - 10 species in database (12/95)

1996

- Set allocations from Congress to NAL
- Move ARS funding to genomic researchers

PGDIC Personnel

Joseph Howard

Director, National Agricultural Library

Public Services Division
Keith Russell, Associate Director
Leslie Kulp, Head, Reference and User Services Branch
Susan McCarthy, Coordinator, PGDIC
Barbara Buchanan, Librarian
Terrance Henrichs, Office Automation Assistant
David Goodman, Office Assistant
Nalini Basavaraj, University of Maryland Graduate Student
Joanne Meil, University of Maryland Graduate Student

Information Systems Division
Pamela Andre, Associate Director
Gary McCone, Head, Database Administration Branch
Douglas Bigwood, PGD Database Manager
Rose Broome, PGD Systems Administrator
Pamela Mitchell, Systems Analyst
John Krainak, Systems Analyst

Technical Services Division
Idalia Acosta, Head, Cataloging Branch
Shirley Edwards, Head, Indexing Branch
Caroline Early, Head, Acquisitions Branch
Lori Starr, Technical Information Specialist
Win Gelenter, Serials Coordinator
Karl Debus, Monographs Coordinator
Gretchen Kaminski, Indexing Workflow Coordinator
Stanislaw Kosecki, Cataloging Branch

PGDIC's Divisional Services

Division **Service**

PSD

Information Services
- Dissemination
- Information Products
- Program Support
- Reference and Referral

Outreach
- Exhibits
- Presentations

TSD

Collection Development
- Computational Genetics Bibliography
- Maize/Rice/Soybean Genetic Newsletters
- Serials and Monographs ordered

AGRICOLA Enhancements
- Increased Genetics and Molecular Genetic Coverage
- 30% Increase in Author Abstracts
- Molecular Sequence Data/653 Identifier Field
- NLM Collaboration Study in Progress (GenBank Scanning)
- Vocabulary and Thesaurus Work (Genetic Terminology)

ISD

Database Development
- Analysis Phase
- Hardware/Software Purchases
- Technical Committee Convened
- Three-Year Plan Written
- First Database Released (AAtDB)

National and International Networking
- French Human Genome Program
- Japanese Rice Genome Program
- Chinese Rice Genome Program
- Canadian Wheat Mapping Group
- International Triticale Mapping Initiative
- German Central Agency for Agricultural Documentation and Information
- CAB International
- Australian/Mendelian Inheritance in Animals Database
- *Arabidopsis*-Multinational Committee
- European Molecular Biology Laboratory
- John Innes Institute/ United Kingdom

Information Services

Information Dissemination:
The Plant Genome Data and Information Center was established to assist the information needs of the USDA Plant Genome Research Program. The services provided by the Center include administrative support, Program outreach, direct and indirect user support, and the development of a database to sustain the research effort.

Information Products
The Plant Genome Data and Information Center staff have produced a variety of information products for the Program.

Newsletter:
Bigwood, D. (1991) Plant Genome Database-Update. Probe 1(1/2):5-6

McCarthy, S. (1991) Information Superhighway Envisioned-Legislation Pending to Establish National Computer Network. Probe 1(1/2):8-9

McCarthy, S. (1991) French Join the International Human Genome Effort. Probe 1(1/2):10

Bigwood, D. (1991) USDA's Plant Genome Database-Collaborative Efforts Continue. Probe 1(3/4):11-12

McCarthy, S. (1991) Columbus Ends Global Isolation. Probe 1 (3/4):27

McCarthy, S. (1991) "Seeds of Change" Quiz. Probe 1 (3/4):28

McCarthy, S. (1991) USDA's Plant Genome Research Program. ALIN 17 (10):1-6

McCarthy, S. (1991) USDA's Plant Genome Program Looks to the Human Genome Project. ALIN 17 (10):6-7

McCarthy, S. (1992) Japan's Rice Genome Program. Probe 2 (2):14-15

McCarthy, S. (1992) Plant Genome Research Grant Program: First Annual Report - 1991. Probe 2(1):4-5

McCarthy, S. (1992) The Federal Biotechnology Commitment. Probe 2(1):14-15

Bigwood, D. (1992) Plant Genome Database to Release Data in ASN.1 Format. Probe 2(2):7

McCarthy, S. (1992) APINMAP An Asian Medicinal Plants Database. Probe 2(3):28-29

Broome, R. (1992) Vocabulary Control in the Plant Genome Database. Probe 2(3):7-8, 17

McCarthy, S. editor (1991) Probe 1(1/2):1-24

McCarthy, S. editor (1991) Probe 1(3/4):1-32

McCarthy, S. editor (1992) Probe 2(1):1-32

McCarthy, S. editor (1992) Probe 2(2):1-28

McCarthy, S. editor (1992) Probe 2(3):1-32

Quick Bibliographies:

Richardson, D. Plant Genome Analysis Techniques: Electroporation Methods and Applications. 150 Citations. 47 pp. QB 92-34

Richardson, D. Breeding for Cold Tolerance in Plants. 212 Citations. 49 pp. QB 92-62

Basavaraj, N. Economic Aspects of Agricultural Bio/Technology. 189 Citations. 43 pp. QB 92-60

McCarthy, S. Ethnobotany and Medicinal Plants: January 1990-June 1991. 591 Citations. 107 pp. QB 92-66

McCarthy, S. Ethnobotany and Medicinal Plants: July 1991-July 1992. 546 Citations. 134 pp. QB 93-02

Special Reference Briefs:

McCarthy, S., Buchanan, B. and Richardson, D. Yeast Artificial Chromosomes. In Preparation.

Richardson, D. and Buchanan, B. Chromosomal Painting. In Preparation.

Nucleotide Sequence Listings:

McCarthy, S. and Henrichs, T. 365 *Zea mays* Sequences

McCarthy, S. and Henrichs, T. 193 *Glycine max* Sequences

McCarthy, S. and Henrichs, T. 289 *Triticum aestivum* Sequences

McCarthy, S. and Henrichs, T. 243 *Arabidopsis thaliana* Sequences

McCarthy, S. and Henrichs, T. 270 *Oryza sativa* Sequences

Press Releases:

McCarthy, S. and Norris, B. Plant Genome I: June 1992

McCarthy, S., Cherry, M. and Cartinhour, S. AAtDB Available for the Public: In final review

Video Tapes:

Bottino, P. Chromosomes and Cell Division. 79 min

Bottino, P. Mendelian Genetics, Linkage and Mapping. 79 min

Bottino, P. Nucleic Acids. 90 min

Bottino, P. DNA Technology. 190 min

Bottino, P. RFLP Technology. 90 min

Probe

Probe is the official publication of the USDA Plant Genome Research Program, and is published quarterly. Volume 1 was published as two combined issues. Volume 2 will be published as four separate issues. Altogether four issues were printed in Fiscal Year 1992.

The mailing list for ***Probe***, was built using a selection of unsolicited mailings and announcements. The following organizations provided mailing addresses or announcements:

- American Society of Plant Physiologists
- Crop Science Society of America
- Maize Genetics Newsletter
- Soybean Genetics Newsletter
- International Triticale Mapping Initiative
- David Neale/ Forest Tree Geneticists
- BIONET: Bionews Group Announcement
- USDA Biotechnology Notes
- Human Genome Newsletter
- ARS Directory
- Biodiversity Electronic Network

Probe is mailed to an international audience. This audience includes legislators, journalists, administrators, and scientists. The permanent list registers nearly 3,500 subscribers. In Fiscal Year 1993, announcements describing the newsletter will be sent to well over 20 associations and professional societies.

Distribution Statistics

Issue	Distributed
Spring Volume 1 number 1/2	8,200
Winter Volume 1 number 3/4	5,900
Spring Volume 2 number 1	10,000
Summer Volume 2 number 2	3,500
Total	***27,600***

Program Support

The Plant Genome Data and Information Center has provided a wide range of administrative and technical support for the USDA Plant Genome Research Program. The Center has organized a number of meetings for the Program including:

- quarterly database design conferences
- Technical Committee
- set agendas and hosted many meetings for Program contacts

The Center has actively developed outside contacts for the Program. Notably, the Director and staff of the French Human Genome Program met with Dr. Jerome Miksche and PGDIC staff.

The Plant Genome Data and Information Center has also provided administrative support to the Program. The Center is managing a Specific Cooperative Agreement with Massachusetts General Hospital and Dr. Howard Goodman. The purpose of the agreement is to develop and deploy a database for the *Arabidopsis* research community.

The Plant Genome Data and Information Center staff have also attended meetings for the Director of the Program at his request.

Specific Cooperative Agreements	Cumulative Total
University of Maryland Dr. Paul Bottino Agreement # 58-0520-1-118	$379,050
University of Maryland Dr. James Reveal Agreement # 58-0520-2-135	$ 50,000
Massachusetts General Hospital Dr. Howard Goodman Agreement # 58-0520-1-150	$300,000
Cornell University Dr. Steven Tanksley Pending	$ 59,950

Reference and Referral Services

The Plant Genome Data and Information Center staff have provided reference and referral services to a diverse audience. Services have been provided to the following:

- Program administrators
- Congressional request
- Start-up entrepreneurial biotechnology company
- Established biotechnology companies
- Academic biotechnology centers
- Government scientists
- University scientists
- General public
- Plant Breeders

Reference services have also been provided to international clientele including: People's Republic of China, Japan, Puerto Rico, the former Soviet Union and others.

PGDIC Reference Requests

Month	FY 1991	FY 1992	
October		43	
November		26	
December		29	
January		10	*1991 Monthly Avg*
February	1	49	5.6
March	5	97	
April	2	8	*1992 Monthly Avg*
May	10	10	28.5
June	1	8	
July	3	9	
August	18	30	
September	5	23	
Total	45	342	

PGDIC Information Product Distribution

Month	FY 1991	FY 1992	
October		8213	
November		1119	
December		364	
January		417	
February	23	1045	*1991 Monthly Avg*
March	1	5301	203.3
April	103	122	
May	200	1024	
June	134	1600	*1992 Monthly Avg*
July	827	372	2,504
August	159	9766	
September	180	706	
Total	1,627	30,049	

Outreach

The Plant Genome Data and Information Center provides outreach services for the USDA Plant Genome Research Program. PGDIC has exhibited at scientific meetings, presented lectures and workshops. Audiences reached include: government, university and industrial scientists; librarians; and administrators.

Presentations

4/19/91 S. McCarthy gave a lecture as invited speaker on the USDA Plant Genome Research Program. University of Oregon. Corvallis, Oregon

5/15/91 S. McCarthy presented to the USDA Animal Genome Planning Committee how NAL could best serve their database needs. Beltsville, Maryland

5/17/91 S. McCarthy gave a lecture as invited speaker on the USDA Plant Genome Research Program. ARS National Center for Agricultural Utilization Research. Peoria, Illinois

5/20/91 S. McCarthy presented NAL's role in developing the Plant Genome Database to the Soybean Advisory Group. St. Louis, Missouri.

9/12/91 S. McCarthy introduced the NAL staff to the Plant Genome Research Program and the PGDIC. Beltsville, Maryland

6/6/92 S. McCarthy gave an Introduction to Biochemistry in 15 minutes for the Intelligenetics Training Workshop. Mountain View, California

6/8/92 S. McCarthy gave a lecture as invited speaker on the USDA's Plant Genome Research Program for the annual meeting of the Special Libraries Association meeting. San Francisco, California

6/30/92 S. McCarthy gave a short introduction to the reference and referral services and tools used in the Center to a University of Maryland Library School graduate class. Beltsville, Maryland

9/24/92 S. McCarthy gave a brief introduction to the PGDIC program and activities to the ARS field librarians and their library committee members. Beltsville, Maryland

Exhibits

7/15-16/91 S. McCarthy and D. Richardson exhibited at the Mid-Atlantic Plant Molecular Biology annual meeting. Baltimore, Maryland

7/29-8/1/91 S. McCarthy and K. Schneider exhibited at the American Society of Plant Physiology annual meeting. Albuquerque, New Mexico

8/21/91 S. McCarthy exhibited at the Capitol Area Biotechnology Information Network meeting. Beltsville, Maryland

9/18/91 S. McCarthy exhibited at the IAALD Conference. Beltsville, Maryland

10/5-12/91 S. McCarthy, T. Henrichs, and L. Starr exhibited at the Third International Congress of Plant Molecular Biology. Tucson, Arizona

3/19-20/92 S. McCarthy exhibited at the Maize Genetics Conference. Asilomar, California

4/21-present PGDIC staff organized the design and implementation of a display in the lobby of the National Agricultural Library (Between April and September over 500 pieces of literature have been distributed from the display area). Beltsville, Maryland

6/21-24/92 S. McCarthy and B. Buchanan exhibited at the World Congress on Cell and Tissue Culture. Crystal City, Virginia

6/28-30/92 S. McCarthy, T. Henrichs and B. Buchanan exhibited at the In Vitro Culture and Horticulture Breeding Conference. Baltimore, Maryland

7/27-29/92 S. McCarthy exhibited at the 4th Biennial Conference on Molecular and Cellular Biology of the Soybean. Ames, Iowa

7/27-28/92 T. Henrichs and B. Buchanan exhibited at the 9th Annual Mid-Atlantic Plant Molecular Biology Society meeting. Beltsville, Maryland

8/1-5/92 S. McCarthy, T. Henrichs and S. Cartinhour exhibited at Annual Meeting for the American Society of Plant Physiologists. Pittsburgh, Pennsylvania

8/10-12/92 S. McCarthy exhibited at the Annual Meeting for the American Institute of Biological Sciences. Honolulu, Hawaii (Partially supported with outside funds)

Workshops

9/10-11/91 PGDIC sponsored a training session for librarians to learn searching the Intelligenetics Software Suite and the databanks, GenBank, PIR, etc. Washington, D.C.

9/12-13/91 PGDIC sponsored a training session for scientists to learn searching the Intelligenetics Software Suite and the databanks, GenBank, PIR, etc. Washington, D.C.

6/6/92 PGDIC sponsored a training session for librarians to learn searching the time-share Intelligenetics Software Suite. This workshop was held in conjunction with a major Genome Program offered at the annual Special Libraries Association meeting. Mountain View, California

Collection Development

Subject Coverage

All aspects of plant and animal genome mapping, including: nucleotide and protein sequencing; automated sequencing and large scale mapping efforts; physical and cytogenetic maps; plant breeding efforts based on or making use of, mapping efforts; research procedures and equipment of potential applicability; scientific and bibliographic databases and software; developments in computational genetics; research that could lead to breakthroughs in collection, analysis, and management of genome data; trends in informatics; and progress of other related efforts/programs such as the Human Genome mapping program.

Computational Genetics Bibliography

Resources have been provided to the National Agricultural Library for improving access to research results and relevant theories. Primary access is achieved through the published literature. PGDIC staff have studied the April 27, 1990 draft version of Computational Genetics Bibliography produced by Sarah Barron. Frequency distribution of citations were compiled and high frequency journals were then acquired. Low frequency sources were acquired on a per-issue basis where ever possible.

The frequency distribution for the serial citations are ranked for the top 21 journals. These journals accounted for 71% of all the serial citations in the April 27, 1990 bibliography.

1 Nucleic Acids Research
2 Bulletin of Mathematical Biology
3 Computer Applications in the Biosciences
4 Proceedings National Academy of Sciences USA
5 Journal of Molecular Biology
6 Journal of Theoretical Biology
7 Science
8 Journal of Molecular Graphics
9 Nature
10 Journal of Molecular Evolution
11 Biochimia Biophysica et Acta

12 Journal of Biomolecular Structure and Dynamics
13 Journal of Chemical Information and Computer Science
14 Biochimie
15 Journal of Applied Mathematics
16 Biochemistry
17 Gene
18 Annual Review of Biophysics and Biophysical Chemistry
19 Biopolymers
20 Mathematical Bioscience
21 Genetic Analysis and Technical Applications

Maize/Rice/Soybean Genetics Newsletters

Similar studies are underway using the last two years of the Maize Genetics Newsletter bibliography (over 1400 citations); the Rice Genetics Newsletter; and the Soybean Genetics Newsletter. The results of these analyses will be used for future materials selection.

Serials and Monographs Ordered

Breakdown of the allocations for monographs and serials are as follows:

Material Type	FY 1991	FY 1992
Monographs	$8,010	$12,756
Serials	$28,553	$68,166
Total	*$36,563*	*$80,922*

Titles Cataloged and Processed in All Formats

FY 91	150
FY 92	306

See Appendix A for a listing of the Plant Genome serial titles.

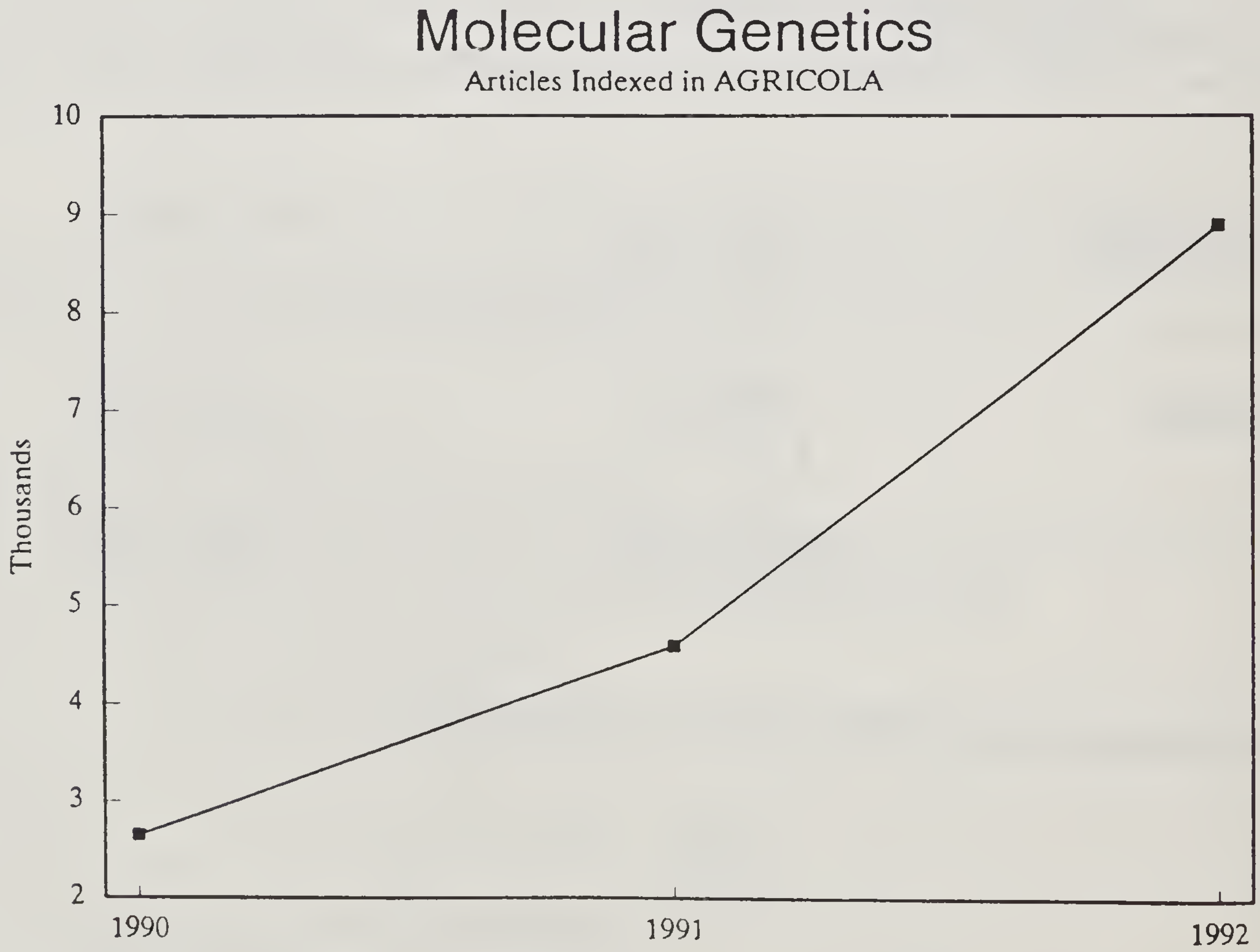
AGRICOLA Enhancements

AGRICOLA is the bibliographic database produced by the National Agricultural Library. AGRICOLA's core subject focus is agriculture, including: production, economics, and nutrition. Related but less fully covered subject areas include: biology, biotechnology, computer science, and information science.

AGRICOLA will provide the primary bibliographic link for the Plant Genome Research Program Database. Resources have been allocated to improve the depth of subject coverage, to expand the scope of subject coverage, and to enhance citations.

Increased Genetics and Molecular Genetics Coverage

In 1990, the baseline year, 2,652 records were added to AGRICOLA related to the field of molecular genetics. In 1992, the second year of the Plant Genome Data and Information Center operation, it is projected that 8,617 molecular genetics related citations will be added to the database. This represents a 3.2 fold increase in subject coverage.



Plant Genome journals recently added for AGRICOLA database indexing include (at least selectively) the following journals:

Nucleic Acids Research
Plant Molecular Biology
Computer Applications in the Biosciences
Journal of Theoretical Biology
Journal of Molecular Evolution
Journal of Biomolecular Structure and Dynamics
Biochimie
Gene
Biopolymers
Conservation Biology

AGRICOLA will have continuous indexing for *Plant Molecular Biology* when the retrospective indexing has been completed. Additional retrospective indexing is underway for the journals listed above, the indexing will cover the last 6 years.

Special indexing requests are also processed for AGRICOLA. The source of these requests include: the Computational Genetics Bibliography, and records needed for the Soybean database. Twelve hundred special indexing requests were sent forward in FY 1992, 51.8% of these requests have now been processed.

Author Abstracts Added to AGRICOLA Records

Author abstracts enhance AGRICOLA records in two ways they improve retrieval and assist the end-user in making literature selections. PGDIC provided funding in FY 1991 to purchase a scanning unit and salary support for technicians to run the scanner. This has dramatically increased the number of AGRICOLA records containing abstracts.

The following indexed journals known to report sequence data are now including author abstracts.

Current Genetics
Developmental Genetics
EMBO Journal
Genome
Journal of Biochemistry
Journal of Biological Chemistry

Journal of Cell Biology
Journal of Molecular Biology
Molecular and Biochemical Parasitology
Molecular Biology
Molecular and General Genetics
Molecular Microbiology
Nature
Nucleic Acids Research
Plant Molecular Biology
Plant Physiology
Plasmid
Proceedings of the National Academy of Science
Science
Soviet Journal of Bioorganic Chemistry
The Plant Cell
UCLA Symposia on Molecular Cell Biology

AGRICOLA Indexing Summary
FY 1990 - FY 1992

Subject Category	FY90	FY91	FY92	Total
Plant Breeding	6156	6306	9000	21,462
% with abstracts	*19.7*	*31.6*	*48.0*	
Molecular Genetics	2652	4587	8890	16,129
% with abstracts	*22.5*	*46.8*	*67.9*	
Economic Crops[1]	837	1589	3065	5,491
% with abstracts	*20.5*	*51.0*	*68.6*	

[1] Economic Crops include: wheat, loblolly pine, soybean, and corn. These four groups were emphasized by ARS in FY 1991-2.

Molecular Sequence Data/ 653 Identifier Field

Molecular Sequence Data is a tag used to identify articles reporting actual sequences. The type of sequence reported is further described in the descriptor field as either a nucleotide or an amino acid sequence. AGRICOLA began using these indexing terms in October 1990.

AGRICOLA Indexing Summary by Identifier/Descriptor
FY 1990 - FY 1992

Subject Category	**FY90**	**FY91**	**FY92**	**Total**
Molecular Sequence Data	4	1210	2232	3,446
% with abstracts	*25.0*	*56.2*	*81.6*	
Nucleotide Sequences	654	1286	2787	4,727
% with abstracts	*17.6*	*57.3*	*79.3*	
Amino Acid Sequences	77	999	21498	3,225
% with abstracts	*24.7*	*57.2*	*79.8*	

NLM Collaboration Study (GenBank Scanning)

Molecular Sequence Data tag is also being used by the National Library of Medicine (NLM) in their program to screen the literature for the national sequence database GenBank. Literature covered in the AGRICOLA database differs from the literature indexed by NLM for MEDLINE (NLM's bibliographic database).

PGDIC staff have identified molecular sequence data records and eliminated those records routinely scanned by NLM. Over 200 records have been sent to NLM for evaluation. Some of the records are known to contain sequences not already found in GenBank, notably insect sequences. These sequences are important to the USDA National Genetics Resources Program which in the near future will expand the Plant Genome Database beyond plant species to include animals, insects, and microorganisms.

See Appendix B for Notes to Indexers Number 21.

Vocabulary and Thesaurus Work (Genetics Terminology)

Vocabulary control in databases aids end user retrieval of information. The CAB Thesaurus is the controlled indexing language of the AGRICOLA database. Terminology used in a rapidly evolving science such as genetics must also evolve. The NAL Thesaurus Management Staff is working to improve the genetics terminology in the CAB Thesaurus. This vocabulary development considers terminological and/or hierarchical compatibility with NLM's Medical Subject Headings. NAL, as one of the three major agricultural databases in the world, is participating with CAB International and United Nations Food and Agriculture Organization to standardize agricultural terminology with the eventual aim of a Unified Agricultural Thesaurus. Draft revisions of genetics, breeding and genetic engineering terminology prepared by NAL Thesaurus Management Staff have been reviewed by this international forum. It is expected that this work will be integrated into the next edition of the CAB Thesaurus, anticipated in 1994.

Vocabulary control for the Plant Genome Database is proceeding simultaneously with database development. An overview of general vocabulary control issues was presented by Lori Starr, NAL Thesaurus Management Staff, to the USDA Plant Genome prototype database developers. Rose Broome, who is heading the effort, has also held a session on vocabulary control for specific database fields of the Plant Genome Database. When appropriate, Rose has applied CAB Thesaurus terminology to the needs of the database. Development of the database vocabulary is being conducted in close collaboration with the USDA Plant Genome prototype database developers.

International efforts to coordinate names given to sequenced plant genes is proceeding under the direction of the International Society of Plant Molecular Biology. Susan McCarthy has been selected to serve as the NAL liaison to the Commission. Commission Chair Carl Price has also been involved in the NAL Plant Genome database design conferences.

Database Development

Analysis Phase

PGDIC consulted with various organizations, industries, and individuals involved in the field of genome informatics. The visits and meetings included the following:

- E.I. du Pont de Nemours and Company
- Agrigenetics
- Johns Hopkins University (Genome Data Base)
- Lawrence Livermore Laboratory
- Lawrence Berkeley Laboratory
- Los Alamos National Laboratory
- European Molecular Biology Laboratory
- John Innes Institute
- Jackson Laboratories (Mouse Genome Database)
- National Center for Biotechnology Information

PGDIC staff are working closely with the five prototype database developers. Site visits have been made to all five laboratories. NAL staff have been present at almost every major meeting held by the principal investigators.

Emerging genome informatics standards indicated that a relational database using the commercial Sybase database management system was the system of choice. This choice should allow the transfer of tools and resources developed for the Human Genome Project to the USDA Plant Genome Research Program. UNIX workstations were selected which are compatible with Sybase.

Internet a national electronic telecommunications network has been identified as a significant access mode for the developing database. NAL has established in FY 1992 its own node to the network with the domain name "nalusda.gov".

The design for the core of the database has been completed. A fully-functional forms-based retrieval system has been written that will allow widespread access to the data. Significant progress has been made towards loading data from the five species groups. Nearly 200,000 records from the AGRICOLA bibliographic database have also been added to the Plant Genome Database (PGD). NAL is actively loading more data into PGD and expect to make the database available to end users by the end of 1993.

Hardware/ Software Purchases

Significant capital investments have been made by way of hardware and software procurements for the Program.

	FY 1991	FY 1992	FY 1993
Hardware	$120,340	$ 24,949	$ 1,931
Software	$ 29,568	$ 13,315	
Total	*$149,908*	*$ 38,264*	*$ 1,931*

See Appendix C for procurement details.

Staff Training

In addition to the capital investments made for hardware and software, PGDIC has committed significant resources in staff training. Training received included advanced Sybase programming and management; and UNIX systems administration.

Database Design Conferences

NAL has hosted six database design conferences with the species groups. Listed below are the dates and locations for each meeting:

April 11, 1991	Madison, WI
July 11, 1991	Beltsville, MD
October 10, 1991	Tucson, AZ
January 23, 1992	St. Louis, MO
April 22, 1992	Beltsville, MD
August 17-18, 1992	Albany, CA

Technical Committee Convened

Experts in database design, implementation, and genetics were gathered in July 1991 to discuss the nature, scope and direction for the Plant Genome Research Program database.

The distinguished participants included:

Olin Anderson
Mary Berlyn
Rose Broome
Michael Cinkosky
Howard Goodman
Leslie Kulp
John McCarthy
Gary McCone
Jimmie Mowder
Oliver Nelson
Ellen Reardon
Robert Robbins
Randy Shoemaker
Scott Tingey
Dennis Benson
Douglas Bigwood
Vincent Caccese
Machi Dilworth
Stephen Heller
David MacKenzie
Susan McCarthy
Jerome Miksche
David Neale
Carl Price
Deborah Richardson
Keith Russell
Quinn Sinnott
Claudia Weston

See Appendix C for a copy of the meeting summary.

Three Year Plan Written

A three year Project Plan has been developed for the NAL Plant Genome Database System. This plan describes the phased development of the database.

See Appendix D for a copy of the plan.

AAtDB First Database Released

Howard Goodman was funded by ARS through NAL to develop a mapping database for *Arabidopsis*. The Goodman laboratory adapted the Human Genome Project database ACeDB for the *Arabidopsis* research community. The database is configured to run as a stand alone system on UNIX workstations as an X-Windows application. A Macintosh version is expected to be released within 6 months.

In addition to adapting ACeDB, the Goodman team has loaded significant amounts of *Arabidopsis* data including:

- Hauge/Goodman cosmid/YAC physical map (>14,000 clones)
- Genetic markers: classical and RFLP
- Unified genetic map (Goodman/Meyerowitz/Classical)
- Primary F2 mapping database (Meyerowitz/Goodman projects)
- Primary two point recombination data of M. Koornneef
- Strain catalog for Nottingham and Ohio State Stock Centers
- Bibliographic citations (1964-present)
- *Arabidopsis* contacts with addresses and phone numbers
- All *Arabidopsis* DNA GenBank sequences (>300 sequences)
- BLASTX defined amino acid sequence similarities
- REBASE restriction enzyme database maintained by R. Roberts
- Graphical displays of genetic/physical maps and sequences

AAtDB is available over the Internet. The Goodman Laboratory is distributing the database via the anonymous file transfer protocol (FTP). A printed manual is available in limited supplies from the Goodman Laboratory. Over 30 sites have downloaded the database.

Marketing for the database has begun. Two demonstrations (listed below) have been funded by the PGDIC. A press release is in the final review process and will be released widely in the near future.

Demonstration Venues of AAtDB

8/1-5/92 Sam Cartinhour and Susan McCarthy. American Society of Plant Physiologists. Pittsburgh, Pennsylvania

9/26-30/92 Michael Cherry. Genome Sequencing and Analysis Conference IV. Hilton Head, South Carolina

See Appendix E for a Fact Sheet describing the database.

National and International Networking

PGDIC has both initiated and fostered national and international contacts for the Program. These contacts include the following:

- French Human Genome Program: Susan McCarthy initiated and coordinated a meeting between the French Human Genome Program staff including: Dr. Jacques Hanoune, Program Director; Pierre Oudet, Director of Informatics; Dr. Michel Cohen-Solal, Research Director of INSERM; and Dr. Michele Durand, Scientific Attache; and the USDA Plant Genome Research Program staff including: Dr. Jerome Miksche and the NAL PGDIC staff on June 19, 1991. Several other meetings have taken place.
- German Central Agricultural Documentation and Information Program. PGDIC staff have had several meetings with both the Dr. Anton Mangstl, Director; and Dr. Frieder Schmidt, Head Information and Coordination Center for Genetic Resources. Susan McCarthy arranged and coordinated the meetings both for NAL staff and outside agencies.
- International Tricale Mapping Initiative Workshop held in Manhattan, KS (9/28-29/91) was attended by Susan McCarthy. Many contacts were made at the meeting. Douglas Bigwood participated in a database workshop associated with the meeting (9/27/91).
- International Society of Plant Molecular Biology. Carl Price and Ellen Reardon have participated in many conferences related to the Plant Genome Database effort. Susan McCarthy has been named liaison with the Plant Gene Nomenclature Commission sponsored by ISPMB.
- CAB International. PGDIC staff have met with Dr. Peter Scott to discuss possible collaborations. A formal invitation has been extended to the USDA Plant Genome Research Program to cooperate on project involving pea (both genetic and germplasm information).
- *Arabidopsis*-Multinational Coordination Committee: PGDIC staff have met with members of the Multinational Coordinating Committee and Dr. Machi Dilworth, NSF-Lead Agency Chair. The coordination has highlighted database development needs.

- Asian Pacific Information Network on Medicinal and Aromatic Plants. Alice Rillo, Coordinator of APINMAP met with Susan McCarthy to discuss the Plant Genome Research Program and the database development.
- Japanese Rice Genome Research Program. Susan McCarthy coordinated a meetings with Dr. Minobe, Program Director; and Dr. Kitamura, Information Leader with the PGDIC staff and separate meetings with the Germplasm Information Network staff (1/13/92). Further assistance has been extended to the Japanese Program with their Rice Genome Newsletter.
- Australian, Mendelian Inheritance in Animals Database. Susan McCarthy organized a meeting with Dr. Frank Nicholas and Dr. Paul Le Tissier and the NAL PGDIC staff. The Plant Genome database development was covered as well as issues related to controlled vocabulary and the Mendelian Inheritance in Animals database.
- Seoul National University. Susan McCarthy met with Dr. Byung-Dong Kim from the Department of Horticulture and the Interdisciplinary Program in Agriculture, Biotechnology Major. The Plant Genome Research Program goals and objectives were presented to Dr. Kim.
- Chinese Rice Genome Research Program. Susan McCarthy coordinated a meeting with Dr. Guo Fan Hong, Program Director; Dr. Wei Jun, National Center for Biotechnology Development; Dr. Qifa Zhang, State Key Laboratory or Crop Genetic Improvement; Dr. Jerome Miksche and key PGDIC staff (9/17/92). A meeting with the Germplasm Resource Information Network staff was also coordinated. Susan McCarthy has extended further assistance to Dr. Qifa Zhang in providing extensive GenBank sequence searching.

GLOSSARY

ALIN - Agricultural Libraries Information Notes
Newsletter of the National Agricultural Library

ASN.1 - Abstract Syntax Notation 1

APINMAP - Asian Pacific Information Network on Medicinal and Aromatic Plants

AAtDB - An *Arabidopsis thaliana* Database
The first database product released

ARS - Agricultural Research Service
An agency of the U.S. Department of Agriculture

ACeDB - A *Caenorhabditis elegans* Database

BLASTX - Software program to compare nucleic acid sequences against all known protein sequences. Produced by the National Center for Biotechnology Information (NLM).

BIONET - A national telecommunications network

DNA - Deoxyribonucleic Acid

FTP - File Transfer Protocol
A method used for transferring large files across a network

ISPMB - International Society of Plant Molecular Biology

IAALD - International Association of Agricultural Information Specialists

ISD - Information Systems Division (NAL)

MEDLINE - Bibliographic database produced by the National Library of Medicine

NAL and the Plant Research Program

Keith Russell
Associate Director for Public Services Division
National Agricultural Library

This program is an outgrowth of a collaborative planning effort between NAL, the Agricultural Research Service, and the Cooperative State Research Service that began in the summer of 1988. That effort culminated in the joint sponsorship of a Crop and Forest Genome Mapping Conference in Washington, D.C., in December of 1988.

With the report of that conference at hand, Clayton Yeutter, as one of his first acts as the newly appointed Secretary of Agriculture in February 1989, announced there would be a USDA Plant Genome Research Program. The rest of the fiscal year, and fiscal year 1990, were devoted to planning the program; it has been underway since funding was approved in late 1990.

We are excited about the NAL role in this program, and via publications and presentations we will keep you advised of progress and changes.

Plant Genome Activities in TSD

Sarah Thomas
Associate Director for Technical Services Division
National Agricultural Library

The NAL Technical Services Division is performing a variety of tasks in support of the Plant Genome Data and Information Center (PGDIC). These include enriching NAL's collection through the selection, acquisition, and processing of new materials, and enhancing the scope of the AGRICOLA database through abstracts, additional citations, and inclusion of data related to sequencing, and improving access to materials through analysis of genetics terminology and updating thesaurus terms.

NAL -	National Agricultural Library
NLM -	National Library of Medicine
NRI-	National Research Initiative *The USDA Competitive Research Grants Program*
PGD -	Plant Genome Database *Mapping Database for the USDA Plant Genome Research Program*
PGDIC -	Plant Genome Data and Information Center (NAL)
PIR -	Protein Identification Resource *Protein sequence database*
PSD -	Public Services Division (NAL)
REBASE -	Restriction Nuclease Database
RFLP -	Restriction Fragment Length Polymorphism
TSD -	Technical Services Division (NAL)
UCLA -	University of California, Los Angeles
USDA -	United States Department of Agriculture
YAC -	Yeast Artificial Chromosome

Appendices

Appendix A

Budget Summary for the

Plant Genome Data and Information Center

Plant Genome Data and Information Center

Budget Summary

Account	1991 Budget	1991 Actual	1992 Budget	1992 Actual	1993 Budget
PSD:					
Assoc. Director	10,000	10,000	8,000	8,000	10,000
RUSB + OD	309,468	284,472	308,474	308,474	306,474
TSD:					
Assoc. Director	9,500	9,500	91,859	91,859	- 0 -
Indexing	125,578	125,578	48,667	48,667	126,526
Cataloging	48,623	33,623	18,000	10,000	10,000
Acquisitions	32,631	32,631	18,000	9,000	8,000
Acq-Materials	100,000	36,563	55,000	72,000	87,000
ISD:					
DAB	187,133	280,566	415,000	415,000	431,000
LAB	17,067	17,067	25,000	25,000	29,000
OD:	170,000	170,000	112,000	112,000	112,000
Total:	1,000,000	1,000,000	1,100,000	1,100,000	1,120,000

Appendix B

Serial and Monograph Titles Ordered in FY 1992

United States
Department of
Agriculture

National
Agricultural
Library

Technical
Services
Division

Beltsville, Maryland
20705

April 24th, 1992

TO: Caroline Early
Head, Acquisitions & Serials Branch

FROM: Karl Debus
Leader, Order/Fiscal Unit

SUBJECT: Title list of Plant Genome Orders for FY 1992

Attached is a list of all titles received and outstanding for the Plant Genome Information Center this fiscal year as of March 31st, 1992. The list is divided into two sections, "Expended" titles (those titles received and already paid for this fiscal year), and "Encumbered" titles (those titles that are outstanding, and, if received, will be paid for out of this year's funds).

The two sections are further subdivided by the 3 Plant Genome fund codes: PG (Plant Genome titles designated for the stacks), PGWT (Plant Genome working tools), and PGBKO (Plant Genome back ordered serials from the Computational Genetics bibliography).

Some titles on the list have their amounts listed as $0.00. These were titles that we expected to pay for, but were received free of charge. Plant Genome gift orders are not included on this list. They are tracked on a different code (GI), and do not require payment of funds.

We have received seven approval books in the Plant Genome subject area. These books total $278.78. These titles are attached on a separate list.

Total (expended and encumbered firm orders through 3/31/92)	=	$21,317.58
Total (approval books)	=	278.78
Total (Shipping - estimate)	=	561.50
Grand total	=	$22,157.86

PLANT GENOME
EXPENDED FUNDS -- FY 1992

FUND ACCOUNT: PG

ORDER NO.	TITLE	AMOUNT
91-0416317	Introduction to artificial intelligence /	$7.79
91-0423302	Genetic algorithms in search, optimization, and machine	$41.08
91-0423312	Software directory for molecular biologists :	$140.40
91-0611321	Statistical methods in biological and medical sciences /	$104.40
91-0611327	Composing user interfaces with InterViews /	$8.00
91-0611406	Statistical analysis of DNA sequence data /	$83.10
91-0628003	Advances in electrophoresis.	$121.80
91-0628209	Genetics and conservation of rare plants /	$43.46
91-0716427	Biotechnology for all.	$10.40
91-0716429	Computer analysis for life sciences /	$120.01
91-0716430	Mathematical models for DNA sequences /	$159.08
91-0716431	Computer applications for molecular biologists /	$48.50
91-0802206	Advances in gene technology :	$46.80
91-0827011	Genetics /	$61.76
91-0827012	Principles of genetics /	$56.88
91-0827013	General genetics /	$51.33
91-0827014	Genetics and molecular biology /	$0.00
91-0827017	Principles of genetics /	$39.98
91-0827018	Genetics /	$42.95
91-0827020	Genetic principles :	$30.59
91-0827021	Cytogenetics :	$120.06
91-0827023	Introduction to genetic analysis.	$39.98

91-0827025	Origins of Mendelism /	$13.01
91-0828001	Viral vectors /	$27.43
91-0828002	Eukaryotic gene expression /	$55.06
91-0828003	Organization and function of the eucaryotic genome :	$18.15
91-0828005	Genetics /	$58.85
91-0828006	Genetic engineering for almost everybody /	$0.00
91-0828007	Primer of genetic analysis :	$32.63
91-0828009	Transposition :	$69.60
91-0828011	Basic genetics /	$53.92
91-0828014	Advances in molecular genetics of the bacteria-plant	$8.20
91-0828015	Microbial genetics	$47.18
91-0828016	Genetics laboratory investigations /	$28.95
91-0829002	Data book of chromosomal aberration test in vitro /	$246.21
91-0829003	Genome multiplication in growth and development :	$73.95
91-0829006	Basic concepts in population, quantitative, and evolutionary	$20.48
91-0829007	Population genetics /	$29.41
91-0829009	Foundations of developmental genetics /	$76.04
91-0829010	Introduction to quantitative genetics /	$43.29
91-0829011	Experiments in gene manipulation /	$37.07
91-0830002	Population genetics and evolution	$89.62
91-0830003	Population genetics and evolution	$45.35
91-0830004	Genetics of populations	$47.18
91-0830006	Correlative neuroanatomy : the anatomical bases of some comm	$19.61
91-0830007	Post-translational modification of proteins by lipids : a la	$37.27

91-0830008	DNA repair	$0.00
91-0830009	Gene structure and mutation : report of symposium held on...	$8.98
91-0904008	Gene regulation : a eukaryotic perspective.	$60.25
91-0904009	Chromosome banding.	$90.00
91-0904010	Site-directed mutagenesis and protein engineering: proceedin	$97.00
91-0904012	Genetic and ecological diversity : the sport of nature.	$26.06
91-0904013	Guidelines and standards for fossil vertebrate databases.	$10.40
91-0904014	Introduction of genetically modified organisms into the envi	$95.17
91-0904015	Conformations and forces in protein folding	$47.23
91-0904862	Biochemical markers in the population genetics of forest tre	$58.17
91-0905009	NMR and biomolecular structure	$77.87
91-0912233	Vocabulaire du genie genetique	$22.18
91-0913002	The human genome project : cracking the genetic code of life	$22.35
91-0916003	Abstracts of papers presented at the 1991 meeting on genome	$20.85
91-0916005	Abstracts of papers presented at the 1991 mtg. on RNA proces	$20.85
91-1003015	Molecular genetics of early Drosophila and mouse development	$20.88
91-1003017	Cell cycle control in eukaryotes /	$17.40
91-1004001	Ubiquitin system /	$21.75
91-1004002	Gap junctions /	$60.90
91-1004003	Genetic maps :	$130.50
91-1004004	Molecular biology of the cytoskeleton /	$53.94
91-1017006	The Faxon Institute 1991 Annual Conference : creating user	$0.00

91-1017417	Studies on the maize Activator (Ac) transposable element in	$55.00
91-1203001	Electrophoresis, supercomputing and the human genome :	$74.82
91-1211001	Bacterial transformation [videorecording]	$30.77
91-1211004	Understanding DNA and gene çloning :	$19.97
91-1211006	evolution of DNA sequences : proceedings of a Royal Society discussion meeting held on 13 and 14 March 1985.	$48.77
91-1211007	DNA cloning/sequencing workshop :	$0.00
91-1211008	Essential molecular biology :	$65.88
91-1212007	DNA sequencing-electrophoresis [VIDEO]	$53.54
91-1212008	Dideoxy DNA sequencing reaction [VIDEO]	$35.92
91-1212009	Preparation of single-stranded DNA templates for sequencing [VIDEO]	$30.75
91-1220001	dictionary of genetics /	$37.55
92-0108012	Behavior and the protein record	$150.26
92-0108213	Fundamentals of plant breeding /	$60.03
92-0109005	Developing and using protein models	$2,397.33
92-0109208	Amphibian cytogenetics and evolution /	$77.86
92-0109209	Genome /	$21.02
92-0115919	Handbook of genetic algorithms /	$45.20
92-0121432	Evolution of a molecular biology software package with applications /	$55.00
92-0128946	Conservation of plant genes : DNA banking and	$0.00
92-0130012	Plant biotechnology 1992 Directory	$30.00
92-0207761	General topology and applications : proceedings of the 1988 Northeast conference /	$93.37
92-0214664	Plant genetic engineering.	$79.56
	TOTAL AMOUNT ACCNT PG :	$6,982.14

FUND ACCOUNT: PGBKO

ORDER NO.	TITLE	AMOUNT
91-0608611	Computers in biology and medicine	$0.00
91-0608613	Computer journal	$8.75
91-0608623	Journal of algorithms	$0.00
91-0608630	Journal of molecular graphics	$0.00
91-0610647	Statistical science	$0.00
91-0611300	Biochemistry abstracts. Part two, Nucleic acids.	$1,899.00
91-0611301	Nucleic Acids Research, Special Publication.	$0.00
91-0620603	Abstracts in biocommerce: ABC	$1,129.69
91-0620604	Genetic analysis, techniques and applications	$8.75
	TOTAL AMOUNT ACCNT PGBKO :	$3,046.19

FUND ACCOUNT: PGWT

ORDER NO.	TITLE	AMOUNT
91-1017007	Genetic maps : locus maps of complex genomes. bk. 6. Plants.	$60.32
91-1119412	Glossary of genetics :	$46.82
91-1120416	Optical publishing directory	$51.33
91-1211003	Protein sequence searching in the REGISTRY	$23.00
92-0130012	Plant biotechnology 1992 directory	$30.00
	TOTAL AMOUNT ACCNT PGWT :	$211.47

TOTAL : $10,239.80

PLANT GENOME
ENCUMBERED FUNDS -- FY 1992

FUND ACCOUNT: PG

ORDER NO.	TITLE	AMOUNT
91-0104017	Human gene mapping techniques :	$39.95
91-0416300	Biodiversity, the videotape.	$24.50
91-0416302	Livestock.	$19.95
91-0416303	Fish and shellfish.	$19.95
91-0416307	Methods in protein sequence analysis /	$89.50
91-0416309	Sciences sources 1991.	$12.00
91-0420325	Neural information processing systems :	$50.00
91-0420329	Distance geometry and conformational calculations /	$50.00
91-0423300	Time warps, string edits, and macromolecules : the theory and practice of sequence comparison	$50.00
91-0423308	1985 Chapel Hill Conference on Very Large Scale Integration	$50.00
91-0423314	A long-range plan for the multinational coordinated Arabidop	$100.00
91-0429322	Chemical applications of topology and graph theory :	$50.00
91-0610300	Homologous recombination and gene targeting /	$34.95
91-0611316	Efficient parallel implementation of sequence analysis	$50.00
91-0611318	Parallel processing of biological sequence comparison	$50.00
91-0611319	Market and economic impact study /	$100.00
91-0611322	DNA probes :	$50.00
91-0611323	Molecular genetics of common metabolic disease /	$50.00
91-0611405	Computer graphics and molecular modeling /	$50.00
91-0620407	Discovery, the search for DNA's secrets /	$50.00

91-0628212	Review of advances in plant biotechnology, 1985-88 :	$50.00
91-0716428	Directory of protein sequence and nucleic acid data sources.	$50.00
91-0720200	Evaluation of intermediate wheatgrass germplasm :	$50.00
91-0802204	Symposium on computer applications in medical care - 1988 /	$50.00
91-0802205	Tools for artificial intelligence workshop :	$50.00
91-0802207	Generalized graphical object editing /	$50.00
91-0827015	Genetic mechanisms /	$25.00
91-0827016	Genes /	$50.00
91-0827019	Genetics /	$50.00
91-0827022	Genetics /	$50.00
91-0827024	Sir Francis Galton and the study of heredity in the	$35.00
91-0827227	molecular basis of imidazolinone herbicide resistance in Ara	$50.00
91-0827228	Genetics of resistance to preharvest aflatoxin accumulation	$50.00
91-0827229	Partial sequence and characterization of a growth hormone	$50.00
91-0828004	Current gene mapping methods /	$25.00
91-0828008	Genetics, society, and decisions /	$50.00
91-0828010	Bacteria, plasmids, and phages :	$50.00
91-0828012	Practical cytology, applied genetics & biostatistics /	$50.00
91-0828013	Understanding genetics /	$50.00
91-0828017	Genetics /	$41.95
91-0829004	Regulation of gene expression :	$50.00
91-0829005	Histone genes : structure, organization, and regulation	$50.00
91-0829008	Genetics and development /	$22.58

91-0830005	Manual of quantitative genetics	$25.00
91-0904011	Longevity, senescence and the genome.	$50.00
91-1003016	Molecular and cellular biology of the yeast Saccaromyces.	$291.00
91-1010004	Entity-relationship approach : the core of conceptual modeli	$50.00
91-1010006	Guidelines on work involving the genetic manipulation of	$25.00
91-1022548	Molecular genetics of common metabolic disease	$1.00
91-1030577	Discovery, the search for DNA's secrets	$1.00
91-1112002	Advances in veterinary science and comparative medicine, 1990.	$50.00
91-1119415	Human genetics	$99.00
91-1127005	maize handbook /	$118.00
91-1211009	DNA sequence knowledge base system (KNOA) /	$50.00
91-1218001	Time warps, string edits, and macromolecules : the theory and practice of sequence comparison	$50.00
91-1224584	Genetics, society, and decisions	$1.00
92-0110008	Cell organelles.	$50.00
92-0114405	Sequence specificity in transcription and translation: proceedings of a conference from a UCLC symposium held in Steamboat Sp	$50.00
92-0114406	Computers in biomedical research	$99.00
92-0115920	General topology I /	$65.00
92-0121200	Journal of computational chemistry	$15.00
92-0121201	IEEE transactions on pattern analysis and machine intelligen	$15.00
92-0121202	Methods of protein and nucleic acid research : immunoelectrophoresis ...	$50.00
92-0207760	GUS protocols : using the GUS gene as a reporter of gene expression /	$30.00
92-0210007	Parallel DNA sequence analysis /	$20.00
92-0210819	plant viruses /	$385.00

92-0214591	Time warps, string edits, and macromolecules : the theory and practice of sequence comparison	$1.00
92-0214595	Genetics /	$1.00
92-0214596	Histone genes : structure, organization, and regulation	$1.00
92-0214597	Directory of protein sequence and nucleic acid data sources.	$1.00
92-0218402	Apomixis in plants /	$129.95
92-0224201	Chemical applications of topology and graph theory :	$50.00
92-0224235	SIAM Review	$50.00
92-0225912	Role of adenosine and adenine nucleotides in the biological system : metabolism, release, transport, receptors,	$267.00
92-0226637	Biochemistry and physiology of polyamines in plants /	$50.00
92-0226638	Plant breeding in the 1990s /	$105.00
92-0226643	Rice biotechnology /	$86.00
92-0226959	Barley: genetics, biochemistry, molecular biology and biotechnology /	$50.00
92-0226960	Ecophysiology of vascular halophytes /	$150.00
92-0226961	Advanced methods in plant breeding and biotechnology /	$50.00
92-0226964	Marine pharmacology : prospects for the 1990s : summary of a California Sea Grant workshop, May 7-9, 1990,	$10.00
92-0304411	Biotechnology guide U.S.A. : companies, data, and analysis /	$199.00
92-0310201	From data banks to data bases : International Conference on the Bacillus Subtilis Genome, Paris, France, September	$50.00
92-0323651	Dictionary of terpenoids /	$1,595.00
92-0324202	Exons, introns, and talking genes : the science behind the Human Genome Project /	$50.00
92-0327617	Genetics and development /	$1.00

92-0327619	Genetics /	$1.00
	TOTAL AMOUNT ACCNT PG :	$6,528.28

FUND ACCOUNT: PGBKO

ORDER NO.	TITLE	AMOUNT
91-0608601	Acta biophysica sinica	$75.00
91-0608602	Advances in applied mathematics	$75.00
91-0608603	Advances in mathematics	$75.00
91-0608604	Advances in applied probability	$75.00
91-0608605	Artificial intelligence	$75.00
91-0608606	Biochimie	$75.00
91-0608607	Biofizika	$75.00
91-0608608	The bulletin of mathematical biophysics	$75.00
91-0608609	CODATA bulletin	$75.00
91-0608610	Cancer letter	$75.00
91-0608612	Computer graphics forum: journal of the European Association	$75.00
91-0608614	Computer methods medicine	$75.00
91-0608615	Computer programs in biomedicine	$75.00
91-0608616	Cray channels	$75.00
91-0608617	EMBO journal	$75.00
91-0608618	IBM journal of research and development	$75.00
91-0608620	IEEE transactions on pattern analysis and machine intelligen	$15.00
91-0608621	IMA journal of mathematics applied in medicine and biology	$75.00
91-0608622	International journal of computer mathematics	$75.00
91-0608624	Journal of applied probability	$75.00
91-0608625	Journal of automated reasoning	$75.00
91-0608626	Journal of combinational theory	$75.00

91-0608627	Journal of computer-aided molecular design	$75.00
91-0608628	Journal of information science	$75.00
91-0608629	Journal of molecular evolution	$75.00
91-0610304	Notices of the American Mathematical Society	$225.00
91-0610305	UCLA Symposia on Molecular and Cellular Biology ; new ser.	$75.00
91-0610631	Journal of the National Cancer Institute	$75.00
91-0610633	CAC/M	$75.00
91-0610635	Lectures on mathematics in the life sciences	$75.00
91-0610636	Mathematical intelligencer	$75.00
91-0610637	The Mathematics scientist	$75.00
91-0610638	Microprocessing and microprogramming	$75.00
91-0610639	Molecular and general genetics	$75.00
91-0610641	Protein sequences & data analysis	$75.00
91-0610643	SIAM journal on discrete mathematics	$75.00
91-0610644	SIAM journal on computing	$75.00
91-0610645	SIAM journal on scientific and statistical computing	$75.00
91-0610646	Rain forest regeneration and management	$75.00
91-0610648	Studies in applied mathematics	$75.00
91-0610649	Journal of computational chemistry	$15.00
91-0610650	Studien zur klassifikation	$75.00
91-0610651	Journal Assoc. Comput. Mach.	$75.00
91-0610652	Linear algebra appl.	$75.00
91-0611302	Protein, nucleic acid and enzyme.	$75.00
91-0611303	Biotechnology Software.	$75.00
91-0620602	Advances in chromatography	$75.00
	TOTAL AMOUNT ACCNT PGBKO :	$3,555.00

FUND ACCOUNT: PGWT

ORDER NO.	TITLE	AMOUNT
91-0410042	Wisconsin package learning guide /	$200.00
91-1119413	Biotechnology research directory : 4000 faculty profiles.	$120.00
91-1220200	User's guide to the software system of the	$70.00
92-0128006	Investing in biotechnology	$20.00
92-0228459	Methods in enzymology	$60.00
92-0302417	1992 GEN guide to biotechnology companies	$407.50
92-0311001	Recombinant DNA technology and applications /	$50.00
92-0311210	Directory of germplasm collections.	$50.00
92-0325400	Roget's international thesaurus.	$17.00
	TOTAL AMOUNT ACCNT PGWT :	$ 994.50
	TOTAL :	$11,077.78

PLANT GENOME APPROVAL BOOKS
FISCAL YEAR 1992
10/1/91 - 3/31/92

Title		Amount
Biotechnologie in der pflanzenzuchtung		$15.47
Plant Cell and Tissue culture		$89.10
Genetically engineered organisms		$43.50
Biosynthesis and the integration . . .		$35.31
Bioelectronics		$30.63
Longevity, senescence, and . . .		$43.46
Human genome project: cracking . . .		$21.31
	Total	$278.78

United States
Department of
Agriculture

National
Agricultural
Library

Technical
Services
Division

Beltsville, Maryland
20705

April 22, 1992

TO: Susan McCarthy
Plant Genome Information Center

Claudia Weston
Technical Services Division

FROM: Caroline Early Ce
Head, Acquisitions & Serials Branch

SUBJECT: Plant Genome Quarterly Report

Library materials allocation:	$55,000
Serials expenditures to date:	$38,107
Monographs expenditures and encumbrances to date	22,158
Library Materials total:	$60,265
Amount remaining:	$-5,265
Overtime used to 4/25:	5,394.61
Amount remaining:	$9,605.39

A detailed list of serials and monographs purchased with these funds is attached.

Plant Genome Journals

$ Cost	Title
631	Abstracts in biocommerce
650	AGRICOLA (2 copies for working tool)
650	AGRICOLA archival discs
200	Bio essays (2nd copy)
5,828	Biochimica et biophysica acta (2nd copy for stacks)
525	Biochemical genetics (2nd copy)
355	Biochimie
54	Biopharm
145	Bio/technology (2nd copy)
2,450	Biotechnology abstracts (CD ROM)
2,995	Biotechnology citation index (CD ROM)
94	Biotechnology software (2nd copy)
552	Bulletin of mathematical biology
689	Comments on molecular and cellular biophysics
162	Comments on theoretical biology
1,125	Computer and information systems abstracts journal
240	Computer applications in the biosciences
440	Computers in biology and medicine (2nd copy)
699	Current genetics
150	Database technology
55	Diversity (WT)
304	DNA and cell biology
200	DNA sequence: journal of DNA sequence and mapping
?	European biotechnology information series
407	GEN Guide to biotechnology companies
85	Gene amplification and analysis
136	Genetic analysis techniques and applications
200	Genetic engineering news: GEN (WT)
?	Genetic maps
130	Genetic resources and crop evaluation
524	Genetica
193	Genome / National Research Council Canada (2nd copy)
66	Genome analysis
310	Genomics
545	Journal of biomolecular structure and dynamics
162	Journal of chemical information and computer sciences
200	Journal of DNA sequencing and mapping
225.	Journal of molecular graphics
60	Life science advances. Molecular genetics
60	Life science advances. Plant physiology
96	Mammalian genome
735	Mathematical biosciences
1,495	MEDLINE (WT)
?	Methods in gene technology
898	Microscopy research and technique
1,750	Molecular & general genetics : MGG (2nd copy)
150	Molecular ecology
350	Nature (2nd copy)
495	Nature genetics
75	New bioresources
	Nucleic acid sequence database (WIP)
1050	Nucleic acids research (2nd copy)
51	Optical publishing directory

Plant Gemone Journals, cont.

115	Plant biotechnology
260	Plant journal for cell and molecular biology
839	Plant molecular biology : an international jour (2nd copy)
1,663	Planta (2nd copy)
380	Proceedings of the National Academy of Sciences (2nd copy)
450	Protein sequences and data analysis
495	Protein science
120	Rice genetics newsletter (2 copies @ 60 ea.)
120	Sci Tech Book news (2 @ 60 ea.)
195	Science (Weekly) (2nd copy)
295	Science watch
58	Scientist (working tool)
405	Trends in biotechnology
350	Trends in cell biology

$35,481 Subtotal + 300? for unknowns + 6.5% serv chg = $38,107 TOTAL

Appendix C

Notes to Indexers

about Molecular Sequence Data

NATIONAL AGRICULTURAL LIBRARY
NOTES TO INDEXERS
No. 21, rev. March 1992, supersedes September 1990

SUBJECT: Molecular Sequence Data

Since December, 1986, AGRICOLA citations have carried the descriptor **NUCLEOTIDE SEQUENCE** for articles discussing this concept whether or not the sequence itself was actually published in the text. Effective immediately, there will be some changes in the manner in which we index articles on molecular sequences.

1) Note that in the 1990 edition of the CAB Thesaurus, the descriptor is in the plural, **NUCLEOTIDE SEQUENCES**, rather than the singular form which AGRICOLA was formerly applying. Continue to apply this descriptor in the descriptor field, field 650, to all articles discussing this concept whether or not the sequence itself is actually published in the text.

2) NAL is proposing a new descriptor to CABI:

   AMINO ACID SEQUENCES
   uf protein sequences

   Definition: the sequence of amino acids within the protein molecule. Amino acid sequences may be deduced from, and often appear in the same article as, nucleotide sequences.

   This descriptor is available for immediate use by AGRICOLA indexers in field 650, and should be applied for articles discussing this concept whether or not the sequence itself is actually published in the text.

In addition to these entries in field 650:

3) If an actual nucleotide or amino acid sequence is printed in the article at hand (which is more often the case than not) or deposited in a databank, add to the identifier field, field 653: MOLECULAR SEQUENCE DATA.

4) If the article references a databank where the sequence data have been deposited, add this information also to field 653 in the format "Databank abbreviation", e.g., SWISSPROT, or "Databank abbreviation/Accession number", e.g., GENBANK/J00207.

Nucleotide sequence accession numbers are usually designated by a single letter followed by a 5-digit number. Be careful to distinguish zeros (0) and "ohs" (O) in the accession number to avoid errors in transcribing. If an accession number is suspect, indexers should photocopy the page on which the accession number occurs and give to the Plant Genome Coordinator. Technicians should input all 653 entries in lowercase even though the examples presented here are in uppercase for emphasis.

Use the following abbreviations for databanks registering molecular sequence data:

DDBJ	DNA Data Bank of Japan
EMBL	EMBL Data Library (nucleotide sequences)
GDB	Genome Data Bank
GENBANK	GenBank Nucleic Acid Sequence Database
HGML	Howard Hughes Medical Inst. Human Gene Mapping Library
OMIM	Online Mendelian Inheritance in Man (McKusick)
PDB	Protein Data Bank (Brookhaven Crystallographic Database)
PIR	Protein Identification Resource (amino acid sequences)
PRFSEQDB	Protein Research Foundation (Amino Acid Seq. Japan)
SWISSPROT	Protein Sequence Database (translated EMBL)

Also, note that some journals are using the abbreviation NBRF or NBRF-PIR. NBRF is an abbreviation for the National Biomedical Research Foundation, the group that sponsors the PIR. For these articles, use only **PIR** as the databank abbreviation. MIPS and JIPID are databanks that work in cooperation with the PIR databank. Do not use **MIPS** or **JIPID** databanks in the 653. Use the **PIR** databank abbreviation for these databanks.

If you encounter a databank which is not on this list, consult the Plant Genome Coordinator immediately.

If sequences are deposited with more than one databank, enter multiple 653's.

If sequences are assigned a joint accession number by GENBANK and EMBL (e.g., GENBANK/EMBL J03482), enter two separate 653's, **GENBANK/J03482; EMBL/J03482**.

Search diligently for this information in the body of the text, in the abstract, in the Materials and Methods section, or in figures. It is often buried in a small-print footnote. This footnote is often on the first page of the article, but sometimes also appears in other places, e.g., with the legend to the figure illustrating the sequence.

Terms entered in the descriptor field, field 650, may be searched with DIALOG suffix code /DE, with BRS label .DE., and on SilverPlatter as **IN DE**.

Terms entered in the identifier field, field 653, may be searched with DIALOG suffix code /ID, with BRS label .ID., and on SilverPlatter as **IN ID**.

Appendix D

Hardware and Software Purchased for the Plant Genome Database System at NAL

INFORMATION SYSTEMS DIVISION

HARDWARE/SOFTWARE PURCHASES

In support of the NAL Plant Genome Database System development, significant capital investments have been made by way of hardware and software procurements. The following charts detail the items purchased, the cost and the fiscal year the procurement was made.

Hardware	FY 1991	FY 1992	FY 1993
Sun Sparcstation 2, configured as fileserver (1)	$25,401		
Server upgrade to Sparc10		$7,950	
Sun Sparcstation 2, (2)	$47,418		
10-port Serial/Parallel terminal server for Sparc		$995	
Sun Sparc printer	$2,500		
External hard disk drives for Sun workstations	$15,836	$3,848	$1,931
Uninterruptible Power Supplies for 5 Sparcstations		$4,670	
Exabyte tape backup device, 5GB, and add-on memory for Sparc's		$6,515	
Tektronix color printer	$8,749		
US Robotics v.32 modems (3)	$1,875		
Ethernet network equipment	$1,528		
Win Personal Computers (3)	$5,706		
Epson printers for PCs (3)	$2,763		
Dell Notebook Computers	$8,564		
CD ROM Drive for PC	$971		
Total	$120,340	$24,949	$1,931

Software	FY 1991	FY 1992
Sybase Database Management Software, consisting of: SQL Server APT Workbench Data Workbench DB-Library Open Client/C	~$17,000	
Transfer Sybase license from ARS to NAL	$1,524	
PC-NFS Networking Software	$632	
Simplify SQL software for Sun	$776	
Windows, Word Perfect, QEMM for 3 PCs	$2,560	
Norton Anti-Virus for 5 PCs		$2,560
LOTUS Upgrade for 1 PC	$476	
CASE Tool: Software Through Pictures for Sparcstation (1)	$6,600	
MARCPlus Conversion Software for Bibliographic data		$3,500
Sybase Software Toolset		$3,851
Corel Draw 3.0 for PC		$368
Lotus Upgrade for PC		$95
Subscription: Silver Platter-Agricola and CAB Abstracts		$2,941
Total	$29,568	$13,315

Appendix E

Summary of the

NAL Technical Committee Meeting

July 10-11, 1991

NAL TECHNICAL COMMITTEE MEETING

National Agricultural Library
Beltsville, Maryland 20705
July 10-11, 1991

NAL's Technical Committee Meeting opened with a description of USDA's Plant Genome Research Program by Jerome (Jerry) Miksche, meeting co-chair and program director. The Program's goal is to locate, transfer, and express genes of economically important crops. Approximately 400 scientists working on over 70 agronomic species at state experiment stations and land grant institutions have received $20-30 million per year since 1988-1989. In FY 91, $14.674 million was allocated to the U.S. Genome Project--$11 million for competitive grants and $3.674 million for a database and its allied support functions (17-20 percent). (Funding for the database ideally should be at 30 percent.)

The species groups chosen to define, establish, and develop relational databases are: Pine (David Neale), Wheat (Olin Anderson), Soybean (Randy Shoemaker), and Maize (Ed Coe). With coordination provided from NAL's database group, the four species groups are to establish a generic relational database across all species in a period of about 6 years. Jerry described activities for each of the 6 years. He closed by indicating that there will be an annual meeting of the competitive grants awardees to present their progress.

David MacKenzie, meeting co-chair, described on-going activities that address future resource needs for biotechnology on a governmentwide basis. Currently there is a "cross cut" study encompassing all Agencies to determine precisely what is going on in biotechnology and how much money is being spent. The study focuses on investments in biotechnological research in the areas of agriculture, the environment, engineering applications, human health, and every possible combination. Agencies are examining their current investments using a broad definition developed by the Office of Technology Assessment in a 1988 report that covers all types of research on the use of life forms for the development of new organisms or products from those organisms. The initiative is slated to gear up in October 1992. The impact of these investments are still being determined. The benefits will need to be studied and tracked.

NAL's Associate Director for Public Services Keith Russell welcomed everyone on behalf of NAL Director Joseph Howard and presented background information on the Library's role in the plant genome program.

Douglas Bigwood, database manager for NAL's Plant Genome Data and Information Center, gave a brief introduction and stated that the focus of the meeting was to derive a set of goals to guide the development of the database (for example, how the finished database should look, what kind of interfaces should exist to external databases, and how users should access the data) and to establish a rough timeframe to accomplish these goals. The proposed approach to the database design and implementation is phased development.

The content of the database and needs analysis are currently being determined by the ARS cooperator prototyping advisory groups and others in the user community. The present working idea is to have the plant mapping data (linkage, physical, and RFLP) as the core subject data. The main mapping activity and the coordination of information between mapping, sequencing, and stock centers will be accomplished at NAL.

It was suggested that an additional area to be added to the agenda was maintenance of the database. The question was raised as to who will provide quality control and continued support.

I. Discussion Points

▪ Before the structure is defined, it should be determined whether the database is intended to be "Observations" or "Truths." Truth is more editorial; observation is not as much work. A layered architecture would provide a stable layer of observation.

▪ A database of maps would be very helpful. One vision of the Plant Genome Research Program is not just maps, but the underlying physiological data as well. When a sequence is put into the database should the old versions be kept, edited, or archived?
Comment: Genbank will keep what authors think is correct.

Comment: There should be three levels in the database: 1) raw data; 2) maps from raw data (interpretive) with public statement from the lab; and 3) official maps. It may not make sense to store raw data in the central database.

▪ Computer readable data does not go through the same classical peer review process as literature. The content should go through some form of quality control. Would the database be able to reject a submission?

II. Discussion on the Database Content

▪ The cooperator prototyping advisory groups are presently conducting or have conducted preliminary needs assessments that indicate:

The **Wheat Group** wants images, every image from every lab.
Initially labs will work on linkage groups.

The **Maize Group** wants maps based on truths and observations. Summarized information can be derived from the raw data. The database can be based on summary only. But that may provide insufficient information for the user.

The **Pine Group** wants to see a database based on summarized information that has been derived from the raw data. Each cross has its own map for segregating populations.

The **Soybean Group** agrees with the other Groups. Tiers of quality information should be in the soybean database. They would like to store data on transient populations as well as permanent populations but on a different level. The research population has indicated that there is a need for the best interpretation with access to raw data. They want images, RFLP probe sequences, and patterns for the five core species.

A comment was made that 20-30 years from now there may not be much need for experimental data but the derived results can be reused in new ways. As a cost consideration, information could be flagged then deleted in the future if not needed. It might be useful to look at some other disciplines like materials properties that have already gone through this process.

▪ Should there just be pointers to the information or should everything be stored? If everything is collected, where will it all be stored?

■ What is NAL's role? If everything is collected, then the data will have to be interpreted. Would NAL interpret the data? It has been proposed that the core groups will rank information then send it to NAL. The Library should publish more than one version of the same map and help develop the tools for database development. Some type of volunteer editorial board would be needed. GenBank and NLM are good examples.

■ Should there be a requirement of publication of the results of research from the Competitive Grants Program? How do you get compliance? The trend now seems to be to build more accountability into the system. Also, there are more voluntary submissions.

III. Discussion on the User Community

■ Categories of users have been identified by the prototyping groups--basic researchers; applied researchers such as post harvest physiologists who try to manipulate genes for engineering; breeders; educators; the general public; and other automated systems/software. There is a need to be able to access other systems in automated procedures. Commercial vendors and developers should be considered. The legal/government community might be a potential user (legislation or patents).

■ Basic and applied research could cause divergence. Tailoring the database to one group would be a mistake; therefore, the lowest common denominator should be used.

■ A scientific database model should reflect reality. Core databases should be established that reflect cross disciplines. Also, the database should not be designed for a particular output.

■ Is there a clear statement of the purpose of the use of the database? It is still being defined. In both the soybean and wheat databases, the utility is not in summarized map information but on segregation, types of markers, germplasm to construct matings, and location of quantitative traits. *Arabidopsis* data needs the lowest level of electronic data storage. In general it is genetic, comprehensive, archival, and dynamic. The basic researcher would like to see the integrated maps, RFLPs, isozymes, consensus maps and background leading to it, sequences, phenotype, closest neighbors and markers, probes, and sequence tagged sites. How these genes interact in different backgrounds, cDNAs similarities to other plants, and function of the genes are also important.

The prototyping groups have prioritized very specifically what each discipline considers important and what will be undertaken first. Getting the genetic maps, RFLP and classical, into the database first is the highest priority on everyone's list. Nuclear cytoplasmic interactions are a much lower priority. The genetic core would include a description of genetic stocks, sites for mapping, pedigrees of stocks, gene products, and molecular markers.

■ The question was raised as to how much information on genetic stocks should be included. Experience with *E. coli* stocks was presented for discussion. Interface with stock center databases should be addressed.

As core data, genetic stock center information is as important as the maps, but most of this information is not currently available electronically. Stocks to be included should come from the appropriate community. Also, stocks should be preserved for posterity to ensure continuance.

■ Use of generic software for development enhances the ability to respond to future needs to accommodate increased workload. It may be more expensive in the short-term but is cost effective in the long-term to meet the needs of the community. Higher level computer-aided software engineering tools should be used for database development.

■ Realistic goals should be viewed in terms of available funds. Future growth potential should be considered. There is a need to ensure accuracy of entered data.

IV. Discussion on Data Flow and Evaluation

■ The working concept is that initially all work will be funnelled through the species groups. The workflow will be formalized in the future. The Library will not have an evaluation/review function. Perhaps expert systems could be developed in the future for first level review.

V. Discussion on Proprietary Information

■ Confidential and proprietary submissions were defined and discussed. There is a need for a policy statement. At this time the thought is that information for the general public will be contained in the database. However, security and access controls are needed since some classes of data may be restricted for the common good.

VI. Discussion on Interfaces with Other Databases

■ GenInfo: The history of GenInfo Backbone and its relationship to GenBank was presented as well as its structure and content. The database is to be distributed on CD-ROM and through INTERNET. NLM is working with NAL to ensure coverage of plant science entries. Patents are being included in the GenInfo Backbone. "Indexed sequences" are being included to facilitate retrieval of related literature. The need to standardize nomenclature was mentioned. Pre-release CD-ROMs contain information from Medline, PIR, and GenBank, allowing retrieval of bibliographic records and their related sequences.

■ GenBank: A database effort should be designed to be scaleable. Most of the data appears in the database prior to publication. The amount of confidential data is small. Electronic publishing was discussed as it relates to GenBank. Cross-links to other databases are viewed by what is the proper domain for each database. The domain for GenBank is the nucleotide sequences. External sources are relied on for other information by pointing to entries in other databases. A registry system has been proposed to accommodate timelag between databases. Software can be written for user interface between GenBank and other databases. Interfaces should accommodate machines as users. Sybase is being used as a client server interface in the human genome community. The structure should have protective layers so that if Sybase is no longer available effects will be minimal. Agencies should look at combined research efforts to find common interfaces. The Library should have a satellite setup of sequence databases with a standard language format like ASN1 for distribution, an application format interface, and adequate documentation of computer activities.

■ GRIN: GRIN was established as a central repository for plant germplasm to enhance communication and to collect and make available accurate data. Maintenance sites (22), which are responsible for entering data, drive the direction of the database. Agronomic and morphological data are currently accessed on a worldwide basis (data are protected). Genetic stock data on barley and soybean will be entered into GRIN. The current database system will be converted to a new hardware and software platform within the next 2 years. Currently there is no interface with other databases. During the developmental phase there were at least seven to eight designs before three to four prototypes were combined into one database.

There are cooperative efforts with the Soviet Union and India. Plans are underway to cooperate with China. GRIN, which catalogs the accessions of the National Plant Germplasm System, is readily accessible by PC using KERMIT communication software.

Timely submission of data to GRIN is motivated by communication from the Germplasm Office National Project Leader (Henry Shands), who informs project participants what is needed and in what timeframe.

What data not currently contained in GRIN should be considered for inclusion in the Plant Genome database? Detailed information on genotypic diversity and how molecular diversity relates to pedigree is not being captured.

■ AGRICOLA: Although currently available on CD-ROM, DIALOG, and BRS there will be a 1-year pilot project to load AGRICOLA on computers at Clemson University using INTERNET. AGRICOLA covers U.S. data and AGRIS covers foreign data. The Library is responsible for co-maintenance of the CAB Thesaurus. Linking AGRICOLA to the Plant Genome database will not pose a problem.

Perhaps a value added CD-ROM could be developed, like GenInfo, where AGRICOLA/MEDLINE plant bibliographic data could be linked with plant sequence data. CRIS (Current Research in Science) records can also be tied into the database since there is a record for each research project.

VII. Discussion on Interfaces (On-Line Access)

■ Hardware – INTERNET (Dial-up Access): INTERNET will be utilized. The GenInfo Backbone will be experimenting with FTS 2000 800 numbers during the next 2-3 years due to the loss of TELENET. They are encouraging users to buy software to run on their own machines. Potential Plant Genome users should be surveyed for dial-up access.

■ Software: Experiments are being conducted at the Department of Energy with software that will run at 9600 baud to work with Microsoft Windows. ASCII and graphic interfaces represent parallel development efforts. ASCII is a throwaway after 3 years. Perhaps instead of utilizing resources to develop interfaces, ASN1 could be utilized with the development of end use PC software.

■ Should images be stored and, if so, in what format? CD-ROM is the least efficient way to capture images. Images are non-core information. NREN (National Research Education Network) will be able to make very high band-width connections available between points on INTERNET that would allow images to be transmitted in a reasonable amount of time. Cost/benefit analysis should be conducted to determine whether graphics should be included for a first pass interface. Users should be queried again to determine true need.

NAL has been putting images up on INTERNET for Document Delivery for about a year. It takes a considerable amount of time on the telephone line. Autorad images could be graphically represented numerically to lower transmission costs.

VIII. Discussion on Other Networks

■ An E-mail interface (BITNET approach) is critical. GenBank is currently using E-mail. There should be some type of interactive access. FTS 800 telephone numbers could provide this at a relatively low cost. Possible service charges were discussed. A small fee would deter hackers. Access through INTERNET is difficult for foreign users at this time; E-mail would be easier.

Should there be other mechanisms besides E-mail to download data? The applications programming interface ASN1 should be considered to send out data. Output should be parsible in machine-readable form and able to be captured on disk.

IX. Discussion on Tools for End Users

■ The present working philosophy is that not much will be provided in the way of tools for data manipulation. The application program interface will allow users to develop their own tools.

The possibility of developing some set of analytic tools that would be helpful for data selection should be examined. (Something to help focus searches.) There should be some tools for novice users and user support. The interface design should be documented for consistency.

X. Open Discussion

■ Nomenclature: The same gene in different organisms should have the same name to facilitate searching for data. A core name should be able to recover all alleles. This needs to be done before it becomes uncontrollable. As more computer systems become available there will be less resistance to standardizing nomenclature. All names should be conserved (historical and current). There must be some consensus process. A standing subcommittee for nomenclature should be appointed.

There is a need to build in a thesaurus capability. The International Society of Plant Molecular Biologists will be addressing this issue at their October meeting with hopes to establish links with this database effort.

■ Off-Line Database Access; CD-ROM as a distribution medium: The currentness of the data determines off-line access. There is a tradeoff between up-to-date hybrid mode CD-ROM and file servers to give updates. CDs are an economic way to distribute data in a short time period. A set timeframe may not be feasible at the beginning. It should depend upon the amount of new information available. CD-ROM release data in ASN1 format with value added computing should not be taken on immediately. Whether or not the cooperators will be responsible for curation and value added activities needs to be determined.

The International Society for Plant Molecular Biology has ambitions to be responsible for curation and valued added activities for the international community. The Society has discussed the feasibility of setting up an office to deal with all aspects of plant molecular genetics, specifically gene sequences and gene function, which will be totally linked to existing databases.

GenBank is distributed via paper, magnetic tape, and CD-ROM, and is also available on USENET.

CD-ROMs are recommended for data distribution as opposed to magnetic tape due to time and money limitations.

■ Off-Line Database Access; Satellite Nodes: Foreign scientists (Europe and Japan) have expressed an interest. CD-ROM might be the best mode of transmission for satellites. Perhaps satellites should not be dealt with until there is a significant workload. The Competitive Grants Program could be used to encourage INTERNET connections. International collaborators have not yet been decided upon but will be sometime in the future.

XII. Discussion on USENET News Groups

■ USENET News groups, available over the INTERNET, utilize specific software that makes news items read like a distributed electronic bulletin board. There should be public forums for the project because more contact means more ideas. The Plant Genome Data and Information Center Newsletter will be utilized for user education. The Newsletter will have an electronic version in the future.

XIII. Other Topics

■ John McCarthy, LBL, discussed issues and questions that face the Library and each prototype group. Other questions that still need to be addressed are:

● There are various ways to use relational databases. Will the authoritative version of the database definition be kept in the form of SQL statements and relational tables or at a higher level? What sort of interface? Without a good interface not many people will use the database. What sort of windows? What sort of interaction style? To what extent are graphics going to be passive or active? How will the interface be implemented? How will these questions be answered? How is the labor divided?

● Database tools such as ERDRAW help productivity during database development. These tools, although still in the developmental phase, can leverage limited resources. Plant Genome has been reviewing and investigating as many sources of these tool as possible.

● Once the core information has been determined, is there a timeframe for implementation?

● Within a year, the individual groups and the central database group expect to have their prototypes designed and some data loaded. The decision had been made in a previous committee that a multiple group effort would be used because of the diversity of information. The groups plan to have their prototypes ready for evaluation and integration by March 1992.

Appendix F

Project Plan for the NAL Plant Genome Database System

Project Plan

NAL Plant Genome Database System

Prepared by:

United States Department of Agriculture
National Agricultural Library
Information Systems Division

DRAFT May 18, 1992

Table of Contents

1 Introduction

1.1 Objectives of the Plant Genome Database System

The main objective of the Plant Genome Database System (PGD) is to provide the user community with information related to plant genomic maps. An equally important objective is to link this information to closely related data pertaining to germplasm, DNA and protein sequence data, and metabolic pathways. The initial versions of the database will contain data for five taxonomic groups: maize, soybeans, wheat, *Arabidopsis*, and pine. Limiting the database to these groups will allow more rapid development and more flexibility; qualities which are required in order to best meet the needs of potential users of this unprecedented project. Once the database has become a stable product it will be much easier to include data for additional species. In addition, it is hoped that the PGD will set a standard that can be followed by those who wish to collect, evaluate, and make use of plant genome data.

1.2 Scope of Document

This document covers the project plan for the Plant Genome Database System at the USDA, National Agricultural Library, Information Systems Division. A summary fiscal year 1991 activities is provided followed by a plan of work which is divided into fiscal years 1992, 1993, 1994, and 1995 and beyond.

2 Summary of FY '91 Activities

2.1 Analysis

2.1.1 Site visits

The first part of the year was spent performing analysis of the problems associated with developing the Plant Genome Database. NAL staff made site visits to many different organizations which have extensive experience in the field of genomic informatics. Many of these organizations have developed their expertise as the result of their association with the Human Genome Project. These organizations include Genome Data Base at Johns Hopkins, Lawrence Livermore National Laboratories, and Lawrence Berkeley Laboratories. Visits were also paid to the two centers primarily responsible for keeping DNA sequence data, Los Alamos National Laboratories and the European Molecular Biology Laboratory. Organizations which have plant genome information systems were also visited. These include DuPont, Agrigenetics, and the John Innis Institute. A site visit was also made to Jackson Laboratories which houses the Mouse Genome Database.

2.1.2 Meetings with Collaborators

Site visits were also made to each of the laboratories of the five principal investigators for the five species groups: maize, soybeans, wheat, pine, and Arabidopsis. Many of these sites were visited multiple times, and by multiple people. NAL staff were present at almost every major meeting held by the principal investigators.

2.1.3 Other Activities

NAL has been active in participating in other activities which will likely have an impact on the database. NAL has been attending the meetings of two of CODATA's commissions (CODATA is an interdisciplinary Scientific Committee of the International Council of Scientific Unions (ICSU)). These commissions are Biological Macromolecules and Standardized Terminology for Access to Biological Data Banks. NAL is also active in the Gene Commission set up by Carl Price which is responsible for research on gene nomenclature. Finally, NAL has developed a collaboration with the Arabidopsis Stock Center to help ensure data compatibility.

2.1.4 NAL Technical Committee Meeting

The analysis phase came to a conclusion with the NAL Technical Committee meeting in July. Experts in database technology and genetics were assembled to advise NAL on its role in the development of the PGD. The attendees of the day-and-a-half meeting offered many helpful suggestions.

2.1.5 Conclusions

The following conclusions were reached as the result of the analysis phase:

1. The Sybase database management system and Sun workstations/servers are

the predominant software/hardware combination in the genomic informatics milieu. Therefore, NAL has decided to conform to this de facto standard.

2. The database development should proceed in a step-wise fashion. A small, but useful portion of the ultimate database should be implemented and made available to the public. An attempt to implement the entire database at one time would result in lengthy delays and perhaps an unsatisfactory product due to the magnitude, scope, and uniqueness, of the project.

3. Due to the fact that several species groups are involved in the project, a coordinator of activities is necessary to ensure data level compatibility. NAL has taken on that role. To fulfill this role NAL has decided to hold meetings periodically (approximately every 3 months) to discuss issues of database design, data compatibility, and data transfer.

4. Although genomic mapping data will constitute the core of the database, inclusion of many types of related data are crucial to the success of the database. The additional information falls into the areas of germplasm/stock centers, bibliographic information, and metabolic pathways. Links will be made between all areas of the database to allow the user to navigate from area to any other.

5. Ideally, all germplasm information would be accessible from the Germplasm Resources Information Network - ARS' germplasm database system. However, GRIN is undergoing a major redesign which will not be implemented for two years. The PGD will proceed with the inclusion of germplasm information and will provide a link to GRIN when that becomes feasible. Meanwhile, GRIN has agreed to incorporate any design decisions made in the area of germplasm if they are suitable. To this end, NAL and GRIN will maintain a continuing dialogue.

6. Links to existing databases will be made whenever possible. NAL will implement satellite nodes of any relevant databases and provide links from PGD to these databases. Genbank and PIR are two such databases. Furthermore, a subset NAL's Agricola database of bibliographic information will be incorporated into the PGD.

7. The initial implementation of the database should be made available to the widest audience possible. Therefore, NAL will provide a character-based user interface which will make the PGD accessible from everything from dumb terminals to advanced workstations. Additional user interfaces will be developed which will enhance the facility with which many users can access the database.

8. The database should be made available in as many different forms as possible. Initially, access will only be available on-line via the Internet, FTS 2000, and modem dial-up. Other versions will be made available on CD-ROM and in a downloadable form. Periodic releases of the data will be made available in the ASN.1 data description language which will allow users to download the data into their own database systems.

2.2 Development

2.2.1 Hardware/Software Procurements

Significant progress has been made towards acquiring and implementing the technology necessary for the development of the PGD. Five Sun SPARCStation 2's have been installed at NAL. These workstations have been connected with an ethernet local area network. This network has been connected to the Internet. The domain name for NAL is nalusda.gov. The Sybase SQL server has been installed on one of the workstations and the front-end development tools (Application Programming Tool, Data Workbench, and DB-Library) have been installed on a second workstation in order to maximize performance. Four disk drives which have a capacity of 1.3 GB have been installed. One of these will hold the database. A second will have an exact duplicate of the database (a "mirror") which will provide redundancy and performance enhancement. The other two drives will be used for software development and temporary storage of data.

In order to expedite software development, a Computer-Aided Software Engineering (CASE) tool has been acquired; Software through Pictures. This tool allows users to graphically design databases and software structures and it automatically generates source code that can be used with Sybase and C compilers thus allowing a more rapid development cycle. Also, changes to software can be more quickly implemented and built-in facilities automatically check the integrity of the design.

2.2.2 Staffing

Staffing currently consists of a Database Manager, a Computer Specialist, and a Programmer/Analyst. Another Programmer/Analyst will start work in mid-February. All personnel have at least one degree in biological science.

3 1992 Plan

Work for FY'92 is divided into three four-month phases. The work to accomplished in each of these phases is described below.

3.1 Phase I

3.1.1 Design

Design of the database has proceeded, with the concentration of the effort in the areas of bibliographic information and mapping data. These designs will be presented at a design review meeting in January. Designs from the collaborators will also be presented and compared for compatibility. It is likely that a consensus on the design for these two areas will be reached at the meeting or shortly thereafter.

3.1.2 Agricola Conversion to Relational Format

The design for bibliographic information has been used as the basis for a large effort to transfer data from the flat file format used in the Agricola CD-ROM, to the relational database format. This data (a subset of the entire Agricola database - about 135,000 records covering 1970 to present - which contains virtually all pertinent records) will be made available to all collaborators. This task is a difficult one due to the original format of the data and the inadequacy of the relational paradigm for representing textual data; considerable programming to overcome these problems is necessary. However, once completed, this will provide a link from other data in the database to citation data and abstracts.

3.1.3 Networking

Networking has been completed with the assignment of the Internet network node at the library. To take advantage of this an electronic mailing list has been set up at the library so that collaborators need only write messages to list in order to communicate with the rest of the collaborators. The list address is genome@nalusda.gov.

3.2 Phase 2

The work during the second four months will concentrate on developing a functional prototype of the database. In order to achieve a viable prototype, two major tasks will be undertaken: 1. the development of a user interface and 2. the development of a protocol for loading data from the species groups.

3.2.1 User Interface

The user interface will be written with APT, the Application Programming tool, which is provided by Sybase as part of their product. APT interfaces are character-based and will run on the widest variety of terminals, PC's, and workstations. APT was used to develop the interface for GDB.

3.2.2 Data Transfer Protocol

The development of a data transfer protocol will require close collaboration with the species groups in order to maintain data-level compatibility. It is likely that the protocol will be based on Sybase's bulk copy facility. Because this is a facility built into Sybase, it will be the most expedient solution in terms of implementation, and hence will allow the most rapid prototyping of the database. However, a certain level of manual manipulation of the data will be necessary and a longer term solution will be required within 2-3 years (or whenever data for additional species are added to the database). One such solution might involve the creation of a program which will read and write ASN.1 formatted data. Regardless of the final form of the protocol, it will contain procedures for data integrity checking.

3.2.3 Design

The third task for this phase will be to solidify the design for the germplasm area of the database. It is likely that another design review will take place during this phase.

3.3 Phase III

In this phase the thrust of the development will be directed towards moving from a prototype to a public version of the database. The major tasks for this period include beta-testing the database/interface, accumulating additional data, and expanding the database to include germplasm data.

3.3.1 Beta-testing

Beta-testing will be performed by a limited group of people which will be drawn from NAL personnel, collaborators and people they designate. The objectives of the beta-test are to expose bugs in the software, fine-tune the user interface, and to receive comment on the utility of the database. Bug-fixes will be implemented as quickly as possible. Comments on the user interface and database utility will be accumulated for about two months. At that time the necessary changes will be implemented.

3.3.2 Data Loading

Data will be added to the database as it is received from the species groups. This period will also serve to test the viability of the data transfer protocol established in the previous phase.

Assuming the design for the germplasm area of the database is finalized during the previous phase, the database will be extended to include this information and the data transfer protocol augmented to reflect this expansion.

3.3.3 Design

Design work will continue on areas of the database not already implemented. A design review meeting is likely to take place during this phase. This should be the last of the design review meetings, assuming a consensus has been reached on the final design.

4 1993 Plan

The major thrust in '93 will be to turn the prototype database into a released product. Implementation and testing of the core database will have been completed at this time, although additional development may be required for "non-core" areas of the database (such as metabolic pathway data). Other tasks will include expanding the choices of user interfaces and implementing a Genbank satellite node.

4.1 Database Release

Although a database, like all software products, is never complete, sufficient development and testing will have been performed so that the database will be "released" to the public. Modem lines will be installed as well as an FTS 2000 circuit. Internet access has already been established. An extensive public relations campaign will be instituted to garner widespread interest. Bug reports from the public will be handled immediately and there will be a periodic review of user comments. The database and/or user interface will be modified if serious deficiencies are exposed.

4.2 Data Loading

Data will continue to be collected throughout this period and the data transfer protocol will be enhanced in an attempt to further automate data loading. Data integrity checking will also be strengthened.

4.3 User Interface Development

Work will commence on two additional interfaces[1]: an electronic mail interface and a graphical interface. The electronic mail interface will provide users who are unable to access the database interactively (e.g. those with Bitnet, but not Internet, access) to pose queries and receive query results. A graphical interface will increase the ease of access to the data through the use of a pointing device (e.g. mouse) and will allow the presentation of graphical data. An obvious use for this type of interface is for the presentation of graphical representations of maps. This

[1] It is possible that additional interfaces will be implemented earlier if the technology can be acquired elsewhere rather than developed from scratch.

interface will probably be based on the X-windows standard because of its widespread availability for almost every type of computer from PC's to mainframes.

4.4 Genbank Satellite Node

The final task for this year will be to implement a satellite node of Genbank. Although the initial implementation of the PGD will include pointers to Genbank accession numbers, a Genbank node at NAL will allow PGD users to navigate easily through sequence data from mapping data and vica-versa. The addition of Genbank node will be relatively easy as Genbank satellites also use the Sybase database management system.

5 1994 Plan

Once the development of the PGD as an on-line product has been completed, the focus will shift to increasing the availability of the data by offering various versions of the database. Also, an interface to GRIN will be developed and a review of the database design will be undertaken.

5.1 CD-ROM Version

Based on the experiences of Genbank and NCBI, it is desirable to offer a non-interactive version of a database, particularly when the volume of data is large. For this reason, the data in the PGD will be distributed on a high-capacity media form, probably CD-ROM. In order to maximize the utility of the data it is likely that the data will be converted to an ASN.1 format. ASN.1 is a standard meta-language which is used to describe data. Data in this format can, with the appropriate tools, be converted into almost any other format and thus can be used by other databases and other interfaces. The ASN.1-formatted data can also be made available for downloading. NCBI will distribute GenInfo data on CD-ROM in ASN.1 format and has developed public domain tools to manipulate ASN.1-formatted data.

5.2 Establishment of Satellite Nodes

Another mechanism for increasing the accessibility of data which has been successful for Genbank is the establishment of satellite nodes. These are exact copies of the main database which are updated automatically over the Internet. The main advantage is that users which do not have adequate access (mainly users without Internet connections, users with slow Internet connections, or users in remote locations such as Europe and Asia) to the PGD at NAL will be able to make effective use of the database. For example, it is envisioned that at least one satellite node will be established in Europe. Satellite nodes also reduce the demand that might otherwise overwhelm the single database access point at NAL.

5.3 Integration with GRIN System

GRIN plans to have an Oracle version of their database running on their newly acquired Unix system. If it is still desirable at that point, an interface will developed jointly with GRIN so that the two databases can be integrated, Thus, users of the PGD will have access to all of the GRIN germplasm data.

5.4 Design Review

Finally, the design of the database will be reviewed at this time in order to determine if it is still adequate to meet the needs of the users. The science of genomics is, of course, a rapidly advancing field of study. NAL plans on adapting quickly to changes in this discipline in order to maximize the utility of the PGD and avoid obsolescence.

6 Plan for 1995 and Beyond

At this time, the PGD should be stable enough such that considerable effort can be spent on adding the data for other plant species to the database. Due to the uncertainty of the state of the computer industry and of genetics, it is impossible to predict what other changes might be necessary in order to maintain the PGD as a state-of-the-art database. However, based upon recent history, it is possible to make a few general conclusions. It is likely that, due to the anticipated volume of data, that new computer hardware will be needed. It is also possible that new database technology, such as object-oriented databases, will find their way into the mainstream and thus allow NAL to develop a database superior to the one developed with Sybase's relational database management system. New user-interface technology that facilitates user access to data is also likely to be developed by the software industry and NAL will take advantage of any advances in this area; this is considered extremely important due to the complexity of data residing in PGD. One such "natural language" interface might allow users to query the database in an ad hoc manner using "plain" English.

Of course, other, unforeseen advances, might arise at any time. A fundamental part of NAL's blueprint for this program is to be flexible enough to respond to change, whenever and wherever it appears.

Appendix G

Fact Sheet about the

Arabidopsis Genome Database

AAtDB

AAtDB

an *Arabidopsis thaliana* database

The second release of the *Arabidopsis* genome database, AAtDB, is now available.

AAtDB, An *Arabidopsis thaliana* Data Base, uses the generalized genome database software ACEDB, created for the *C. elegans* genome effort by Dr. Richard Durbin (MCR-LMB, UK) and Dr. Jean Thierry-Mieg (CNRS-CRMB, France). AAtDB is funded by the U. S. Department of Agriculture Plant Genome Research Program through the National Agricultural Library and is maintained by a group at the Massachusetts General Hospital and Harvard Medical School.

AAtDB is available free of charge through Internet network transfer using Anonymous FTP from either weeds.mgh.harvard.edu or ncbi.nlm.nih.gov.

ACEDB is unlike most databases in that users interact with it primarily by using a computer mouse rather than by typing commands via the keyboard. Information is presented in windows using both text and graphics. The user finds out more about a topic, or more about related topics, by pointing and clicking with the mouse. A powerful query facility is also available. However, our experience is that most users choose the mouse interface to find the information they are interested in.

Some of the major topics AAtDB currently contains:

- The Hauge/Goodman cosmid/YAC physical map including >14,000 cosmid clones.
- Genetic markers, both RFLP and classical markers.
- Unified Genetic Map. Including both the Goodman, Meyerowitz and Scolnik molecular markers and classical genetic markers collected by Koornneef.
- Primary F2 and RI mapping data from the Goodman, Meyerowitz and Scolnik mapping projects.
- Primary two point recombination data from M. Koornneef.
- A strain catalog including all strains and clones available from the Nottingham Stock Centre and the ABRC at Ohio State University.
- Bibliographic citations from 1964 to present, currently numbering over 2,800.
- List of *Arabidopsis* researchers including mail address, phone number, FAX number and electronic mail address. Currently information on over 600 colleagues is included.
- Green Book. The Green Book by Meyerowitz and Pruitt has been updated and integrated into many parts of the database, including phenotype and allele descriptions.
- All *Arabidopsis* DNA sequences from GenBank, currently there are over 400 sequences.
- BLASTX defined amino acid sequence similarities for all *Arabidopsis* sequences against the SwissProt, PIR and GenPept protein sequence databases.
- REBASE restriction enzyme database maintained by R. Roberts, CSHL.
- Graphical displays of all Genetic Maps, Physical Maps, and DNA Sequence features and similarities.

As much as possible all information is connected to other information in the database. The database presents the information in separate windows that allow many parts of the collection to be viewed at one time. There are also many paths to any topic, allowing the user to easily navigate the connections between the various types of information.

This is just the starting point. Just as the known *Arabidopsis* genome information is always being expanded AAtDB is also being extended and enhanced via periodic updates. There is much more information that we are working to include in subsequent updates of AAtDB. For example scanned images of photographs showing mutant phenotypes, RFLP autoradiograms and RAPD gels will be an optional feature of the database.

The database currently requires a Unix™ workstation running X-Windows. Versions of the ACEDB database software are available for Sun Microsystems SPARCstations™, Digital Equipment's DECstation™, Silicon Graphics Iris™ series and NeXT™ workstations.

A printed manual "An Introduction to ACEDB: For AAtDB, An *Arabidopsis thaliana* database" is available on request from the MGH group. While this manual is not required to use the database many people have found it very useful as a beginning tutorial for the ACEDB software.

A Macintosh™ version of the ACEDB software is under development. We are currently using an beta version of the sofware and are very impressed. A public release version is expected within six months.

All the information in AAtDB is also available using Gopher and WAIS client software. Gopher and WAIS are free software that utilitizes the Internet computer network to query and retrieve information. If you have Gopher or WAIS client software point it at the host weeds.mgh.harvard.edu, for WAIS the database names are AAtDB and Arabidopsis-Biosci.

For more information contact J. Michael Cherry or Samual W. Cartinhour
FAX: 617-726-6893
Electronic Mail: curator@frodo.mgh.harvard.edu

Samual W. Cartinhour, J. Michael Cherry and Howard M. Goodman
Department of Molecular Biology, Massachusetts General Hospital
Department of Genetics, Harvard Medical School
Boston, Massachusetts
USA